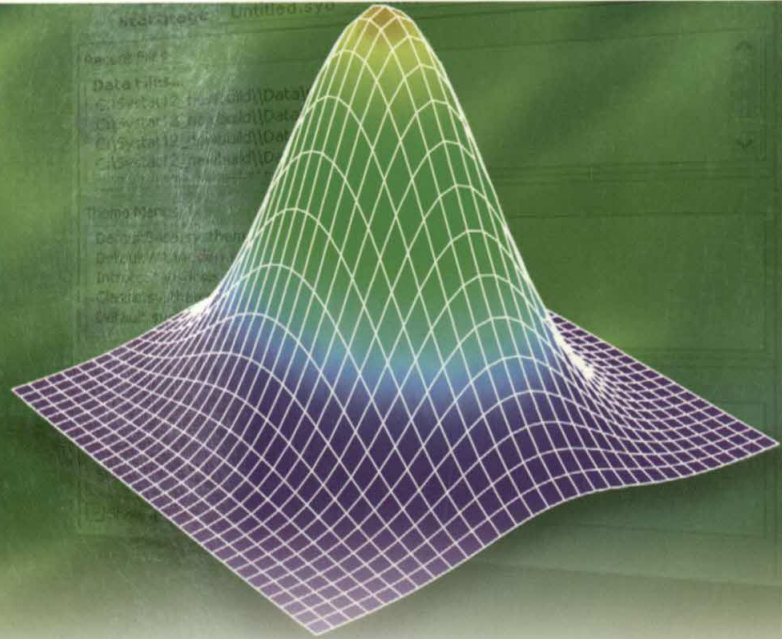


3.29

6

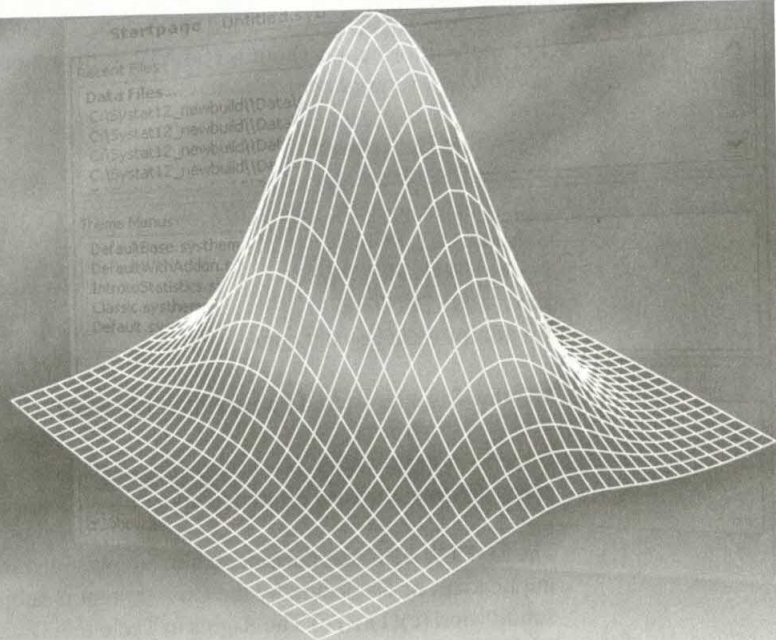
# SYSTAT<sup>®</sup> 12



## Statistics III

~~Dr. Ref~~ Fw Acc  
~~Lib~~ SCERT  
lib  
28/8  
28/8-100

# SYSTAT<sup>®</sup> 12



## Statistics III



**SYSTAT<sup>®</sup>**  
WWW.SYSTAT.COM



For more information about SYSTAT<sup>®</sup> software products, please visit our WWW site at <http://www.systat.com> or contact

Marketing Department  
SYSTAT Software, Inc.  
1735 Technology Dr., Ste. 430  
San Jose, CA 95110  
Phone: (800) 797-7401  
Fax: (800) 797-7406  
Email: [info-usa@systat.com](mailto:info-usa@systat.com)

Windows is a registered trademark of Microsoft Corporation.

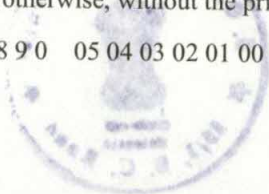
General notice: Other product names mentioned herein are used for identification purposes only and may be trademarks of their respective companies.

The SOFTWARE and documentation are provided with RESTRICTED RIGHTS. Use, duplication, or disclosure by the Government is subject to restrictions as set forth in subdivision (c)(1)(ii) of The Rights in Technical Data and Computer Software clause at 52.227-7013. Contractor/manufacturer is SYSTAT Software, Inc., 1735 Technology Drive, Suite 430, San Jose, CA 95110. USA.

SYSTAT<sup>®</sup> 12 Statistics- III  
Copyright © 2007 by SYSTAT Software, Inc.  
SYSTAT Software, Inc.  
1735 Technology Dr., Ste. 430  
San Jose, CA 95110  
All rights reserved.  
Printed in the United States of America.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, without the prior written permission of the publisher.

1 2 3 4 5 6 7 8 9 0 05 04 03 02 01 00



24.2.2015  
14622

---

# Contents

<i>List of Examples</i>	<i>xxxiii</i>
-------------------------	---------------

## *Statistics I*

<i>1 Introduction to Statistics</i>	<i>I-1</i>
-------------------------------------	------------

Descriptive Statistics . . . . .	I-1
Know Your Batch . . . . .	I-2
Sum, Mean, and Standard Deviation . . . . .	I-3
Stem-and-Leaf Plots . . . . .	I-3
The Median . . . . .	I-4
Sorting . . . . .	I-5
Standardizing . . . . .	I-6
Inferential Statistics. . . . .	I-7
What is a Population? . . . . .	I-7
Picking a Simple Random Sample. . . . .	I-8
Specifying a Model . . . . .	I-10
Estimating a Model . . . . .	I-10
Confidence Intervals. . . . .	I-11
Hypothesis Testing. . . . .	I-12
Checking Assumptions . . . . .	I-14
References . . . . .	I-16

## ***2 Bootstrapping and Sampling***

***I-17***

Statistical Background . . . . .	I-17
Resampling in SYSTAT . . . . .	I-21
Resampling Tab . . . . .	I-21
Using Commands . . . . .	I-22
Usage Considerations . . . . .	I-22
Examples . . . . .	I-23
Computation . . . . .	I-38
Algorithms . . . . .	I-38
Missing Data . . . . .	I-38
References . . . . .	I-39

## ***3 Classification and Regression Trees***

***I-41***

Statistical Background . . . . .	I-42
The Basic Tree Model . . . . .	I-42
Categorical or Quantitative Predictors . . . . .	I-45
Regression Trees . . . . .	I-45
Classification Trees . . . . .	I-46
Stopping Rules, Pruning, and Cross-Validation . . . . .	I-47
Loss Functions . . . . .	I-48
Geometry . . . . .	I-48
Classification and Regression Trees in SYSTAT . . . . .	I-51
Classification and Regression Trees Dialog Box . . . . .	I-51
Using Commands . . . . .	I-54
Usage Considerations . . . . .	I-54
Examples . . . . .	I-54
Computation . . . . .	I-62
Algorithms . . . . .	I-62
Missing Data . . . . .	I-62
References . . . . .	I-62



## **4 Cluster Analysis**

**I-65**

Statistical Background . . . . .	I-66
Types of Clustering . . . . .	I-66
Correlations and Distances . . . . .	I-67
Hierarchical Clustering . . . . .	I-68
Partitioning via K-Clustering . . . . .	I-78
Additive Trees . . . . .	I-80
Cluster Analysis in SYSTAT . . . . .	I-82
Hierarchical Clustering Dialog Box . . . . .	I-82
K-Clustering Dialog Box . . . . .	I-88
Additive Trees Clustering Dialog Box . . . . .	I-91
Using Commands . . . . .	I-93
Usage Considerations . . . . .	I-95
Examples . . . . .	I-96
Computation . . . . .	I-122
Algorithms . . . . .	I-122
Missing Data . . . . .	I-122
References . . . . .	I-122

## **5 Conjoint Analysis**

**I-125**

Statistical Background . . . . .	I-125
Additive Tables . . . . .	I-126
Multiplicative Tables . . . . .	I-128
Computing Table Margins Based on an Additive Model . . . . .	I-130
Applied Conjoint Analysis . . . . .	I-131
Conjoint Analysis in SYSTAT . . . . .	I-133
Conjoint Analysis Dialog Box . . . . .	I-133
Using Commands . . . . .	I-135
Usage Considerations . . . . .	I-135
Examples . . . . .	I-136

Computation . . . . .	I-152
Algorithms . . . . .	I-152
Missing Data . . . . .	I-153
References . . . . .	I-153

## ***6 Correlations, Associations, and Distance Measures*** ***I-157***

Statistical Background . . . . .	I-158
The Scatterplot Matrix (SPLOM) . . . . .	I-159
The Pearson Correlation Coefficient . . . . .	I-160
Other Measures of Association . . . . .	I-161
Transposed Data . . . . .	I-167
Hadi Robust Outlier Detection . . . . .	I-168
Simple Correlations in SYSTAT . . . . .	I-170
Simple Correlations Dialog Box . . . . .	I-170
Using Commands . . . . .	I-177
Usage Considerations . . . . .	I-178
Examples . . . . .	I-179
Computation . . . . .	I-199
Algorithms . . . . .	I-199
Missing Data . . . . .	I-199
References . . . . .	I-200

## ***7 Correspondence Analysis*** ***I-201***

Statistical Background . . . . .	I-201
The Simple Model . . . . .	I-202
The Multiple Model . . . . .	I-203
Correspondence Analysis in SYSTAT . . . . .	I-204
Correspondence Analysis Dialog Box . . . . .	I-204

Smart Correspondence Analysis Dialog Box . . . . .	I-205
Using Commands . . . . .	I-206
Usage Considerations . . . . .	I-206
Examples . . . . .	I-207
Computation . . . . .	I-218
Algorithms . . . . .	I-218
Missing Data . . . . .	I-218
References . . . . .	I-218

## **8 Crosstabulation**

### ***(One-Way, Two-Way, and Multiway) I-219***

Statistical Background . . . . .	I-220
Making Tables . . . . .	I-220
Significance Tests and Measures of Association . . . . .	I-222
Crosstabulations in SYSTAT . . . . .	I-228
One-Way Frequency Tables Dialog Box . . . . .	I-228
Two-Way Tables Dialog Box . . . . .	I-231
Multiway Tables: Tabulate Dialog Box . . . . .	I-237
Using Commands . . . . .	I-244
Usage Considerations . . . . .	I-246
Examples . . . . .	I-248
References . . . . .	I-296

## **9 Descriptive Statistics**

### ***I-297***

Statistical Background . . . . .	I-299
Location . . . . .	I-299
Spread . . . . .	I-301
The Normal Distribution . . . . .	I-301
Test for Normality . . . . .	I-302



Multivariate Normality Assessment . . . . .	I-303
Non-Normal Shape . . . . .	I-303
Subpopulations . . . . .	I-305
Descriptive Statistics in SYSTAT . . . . .	I-307
Basic Statistics Dialog Box . . . . .	I-307
Stem-and-Leaf Plot Dialog Box . . . . .	I-314
Basic Statistics for Rows . . . . .	I-316
Row Stem-and-Leaf Plot Dialog Box . . . . .	I-320
Cronbach's Alpha Dialog Box . . . . .	I-321
Using Commands . . . . .	I-322
Usage Considerations . . . . .	I-323
Examples . . . . .	I-324
Computation . . . . .	I-344
Algorithms . . . . .	I-344
References . . . . .	I-344

## ***10 Design of Experiments***

***I-345***

Statistical Background . . . . .	I-346
The Research Problem . . . . .	I-346
Types of Investigation . . . . .	I-347
The Importance of Having a Strategy . . . . .	I-348
The Role of Experimental Design in Research . . . . .	I-349
Types of Experimental Designs . . . . .	I-349
Factorial Designs . . . . .	<b>I-350</b>
Response Surface Designs . . . . .	I-354
Mixture Designs . . . . .	I-357
Optimal Designs . . . . .	I-362
Choosing a Design . . . . .	I-366
Design of Experiments in SYSTAT . . . . .	I-368
Design of Experiments Wizard . . . . .	I-368
Classic Design of Experiments . . . . .	I-369
Using Commands . . . . .	I-370

Usage Considerations . . . . .	I-370
Examples . . . . .	I-371
References . . . . .	I-388

## ***11 Discriminant Analysis*** ***I-391***

Statistical Background. . . . .	I-392
Linear Discriminant Model. . . . .	I-392
Robust Discriminant Analysis . . . . .	I-399
Discriminant Analysis in SYSTAT . . . . .	I-400
Classical Discriminant Analysis Dialog box . . . . .	I-400
Robust Discriminant Analysis Dialog Box. . . . .	I-405
Using Commands. . . . .	I-407
Usage Considerations . . . . .	I-408
Examples . . . . .	I-409
References . . . . .	I-450

## ***12 Factor Analysis*** ***I-453***

Statistical Background. . . . .	I-453
A Principal Component. . . . .	I-454
Factor Analysis . . . . .	I-457
Principal Components versus Factor Analysis . . . . .	I-460
Applications and Caveats. . . . .	I-461
Factor Analysis in SYSTAT. . . . .	I-462
Factor Analysis Dialog Box . . . . .	I-462
Using Commands. . . . .	I-468
Usage Considerations. . . . .	I-468
Examples . . . . .	I-469
Computation . . . . .	I-492
Algorithms . . . . .	I-492

Missing Data . . . . .	I-492
References . . . . .	I-493

## ***13 Fitting Distributions*** ***I-495***

Statistical Background . . . . .	I-495
Goodness-of-Fit Tests . . . . .	I-496
Fitting Distributions in SYSTAT . . . . .	I-498
Fitting Distributions: Discrete Dialog Box . . . . .	I-498
Fitting Distributions: Continuous Dialog Box . . . . .	I-499
Using Commands . . . . .	I-501
Usage Considerations . . . . .	I-503
Examples . . . . .	I-503
Computation . . . . .	I-518
Algorithms . . . . .	I-518
References . . . . .	I-518

## ***14 Hypothesis Testing*** ***I-519***

Statistical Background . . . . .	I-520
One-Sample Tests and Confidence Intervals for Mean and Proportion	I-520
Two-Sample Tests and Confidence Intervals for Means and Proportions	I-520
Tests for Variances and Confidence Intervals . . . . .	I-521
Tests for Correlations and Confidence Intervals . . . . .	I-522
Multiple Tests . . . . .	I-522
Hypothesis Testing in SYSTAT . . . . .	I-523
Tests for Mean(s) . . . . .	I-523
Tests for Variance(s) . . . . .	I-531
Tests for Correlation(s) . . . . .	I-535
Tests for Proportion(s) . . . . .	I-538



Using Commands . . . . .	I-541
Usage Considerations . . . . .	I-543
Examples . . . . .	I-544
References . . . . .	I-566

# ***Statistics II***

## ***1 Linear Models***

## ***II-1***

Simple Linear Models . . . . .	II-1
Equation for a Line . . . . .	II-2
Least Squares . . . . .	II-5
Estimation and Inference . . . . .	II-5
Standard Errors . . . . .	II-7
Hypothesis Testing . . . . .	II-7
Multiple Correlation . . . . .	II-8
Regression Diagnostics . . . . .	II-9
Multiple Regression . . . . .	II-12
Variable Selection . . . . .	II-15
Using an SSCP, a Covariance, or a Correlation Matrix as Input	II-18
Analysis of Variance . . . . .	II-19
Effects Coding . . . . .	II-20
Means Coding . . . . .	II-21
Models . . . . .	II-22
Hypotheses . . . . .	II-23
Multigroup ANOVA . . . . .	II-24
Factorial ANOVA . . . . .	II-24
Data Screening and Assumptions . . . . .	II-25
Levene Test . . . . .	II-25
Pairwise Mean Comparisons . . . . .	II-26

Linear and Quadratic Contrasts . . . . .	II-28
Repeated Measures . . . . .	II-31
Assumptions in Repeated Measures . . . . .	II-32
Issues in Repeated Measures Analysis . . . . .	II-33
SYSTAT's Sum of Squares . . . . .	II-34
References . . . . .	II-36

## ***2 Linear Models I: Linear Regression II-39***

Linear Regression in SYSTAT . . . . .	II-41
Least Squares Regression Dialog Box . . . . .	II-41
Ridge Regression . . . . .	II-48
Ridge Regression Dialog Box. . . . .	II-49
Bayesian Regression . . . . .	II-50
Bayesian Regression Dialog Box . . . . .	II-51
Using Commands . . . . .	II-53
Usage Considerations . . . . .	II-54
Examples. . . . .	II-55
Computation . . . . .	II-104
Algorithms . . . . .	II-104
References . . . . .	II-104

## ***3 Linear Models II: Analysis of Variance II-107***

Analysis of Variance in SYSTAT . . . . .	II-108
Analysis of Variance: Estimate Model Dialog Box. . . . .	II-108
Analysis of Variance: Hypothesis Test Dialog Box . . . . .	II-113
Analysis of Variance: Pairwise Comparisons Dialog Box . . . . .	II-117
Using Commands . . . . .	II-121
Usage Considerations . . . . .	II-121
Examples. . . . .	II-122

Computation . . . . .	.II-171
Algorithms . . . . .	.II-171
References . . . . .	.II-171

## ***4 Linear Models III: General Linear Models II-175***

General Linear Models in SYSTAT. . . . .	.II-177
Model Estimation (in GLM) . . . . .	.II-177
Hypothesis Test. . . . .	.II-186
Pairwise Comparisons . . . . .	.II-195
Post hoc Tests for Repeated Measures . . . . .	.II-199
Using Commands. . . . .	.II-200
Usage Considerations. . . . .	.II-201
Examples . . . . .	.II-203
Computation . . . . .	.II-249
Algorithms . . . . .	.II-249
References . . . . .	.II-249

## ***5 Introduction to Linear Mixed Models II-251***

Mixed Models and Paired t-test . . . . .	.II-251
Fixed Effects Versus Random Effects . . . . .	.II-255
Why Use Random Effects? . . . . .	.II-259
Some Linear Model Terminology . . . . .	.II-261
String and Numeric Variables . . . . .	.II-261
Estimability . . . . .	.II-262
Data Layout: Multiway or Nested . . . . .	.II-262
Nested Layout . . . . .	.II-266
Balanced and Unbalanced Data. . . . .	.II-267
SYSTAT Notation for Random Effects . . . . .	.II-267
Covariance Structures . . . . .	.II-269



Using Covariates: Regression . . . . .	II-276
Estimation and Prediction . . . . .	II-279
Estimating the Fixed Effects . . . . .	II-279
Estimating Covariance Matrices . . . . .	II-281
Testing Hypotheses . . . . .	II-286
The F Matrix . . . . .	II-287
The D Matrix . . . . .	II-288
The R Matrix . . . . .	II-289
Pairwise Comparison Tests . . . . .	II-290
Diagnostics . . . . .	II-290
Residual Diagnostics . . . . .	II-291
Further Insights	
Henderson's Mixed Model Equation	II-293
Some Properties of BLUPs . . . . .	II-294
Why Random Effect Coefficients are Always Estimable. . . . .	II-295
ML and REML . . . . .	II-295
References . . . . .	II-297

## ***6 Variance Components Models***

***II-299***

Statistical Background . . . . .	II-299
Variance Components in SYSTAT . . . . .	II-301
Model Estimation (in VC) . . . . .	II-301
Hypothesis Test . . . . .	II-306
Using Commands . . . . .	II-310
Usage Considerations . . . . .	II-310
Examples . . . . .	II-311
References . . . . .	II-342

## ***7 Linear Mixed Models***

***II-343***

Statistical Background . . . . .	II-344
----------------------------------	--------

Linear Mixed Models in SYSTAT . . . . .	II-345
Model Estimation (in MIXED) . . . . .	II-345
Category . . . . .	II-347
Random . . . . .	II-348
Options . . . . .	II-350
Hypothesis Tests . . . . .	II-352
F and R Matrices . . . . .	II-354
D Matrix . . . . .	II-355
Using Commands . . . . .	II-356
Usage Considerations . . . . .	II-356
Examples . . . . .	II-357
References . . . . .	II-384

## ***8 Hierarchical Linear Mixed Models*** ***II-385***

Statistical Background. . . . .	II-386
Hierarchical Linear Mixed Models in SYSTAT . . . . .	II-387
Model Estimation (in MIXED) . . . . .	II-387
Hypothesis Test. . . . .	II-394
Using Commands . . . . .	II-398
Usage Considerations. . . . .	II-398
Examples . . . . .	II-399
References . . . . .	II-419

## ***9 Mixed Regression*** ***II-421***

Statistical Background. . . . .	II-422
Historical Approaches . . . . .	II-423
The General Mixed Regression Model . . . . .	II-424
Model Comparisons . . . . .	II-431
Mixed Regression in SYSTAT . . . . .	II-431

Mixed Regression: Hierarchical Data . . . . .	II-431
Data Structure . . . . .	II-438
Using Commands . . . . .	II-441
Usage Considerations . . . . .	II-441
Examples . . . . .	II-442
Computation . . . . .	II-484
Algorithms . . . . .	II-484
References . . . . .	II-485

## ***Statistics III***

### ***1 Logistic Regression***

### ***III-1***

Statistical Background . . . . .	III-2
Binary Logit . . . . .	III-2
Multinomial Logit . . . . .	III-5
Conditional Logit . . . . .	III-5
Discrete Choice Logit . . . . .	III-7
Stepwise Logit . . . . .	III-9
Logistic Regression in SYSTAT . . . . .	III-10
Estimate Model Dialog Box . . . . .	III-10
Quantiles . . . . .	III-18
Simulation . . . . .	III-19
Hypothesis . . . . .	III-20
Using Commands . . . . .	III-22
Usage Considerations . . . . .	III-22
Examples . . . . .	III-24
Computation . . . . .	III-85
Algorithms . . . . .	III-85
Missing Data . . . . .	III-86

References . . . . .	III-89
----------------------	--------

## **2 Loglinear Models** **III-93**

Statistical Background. . . . .	III-94
Fitting a Loglinear Model . . . . .	III-95
Loglinear Models in SYSTAT . . . . .	III-96
Loglinear Model: Estimate Dialog Box . . . . .	III-96
Frequency Table (Tabulate) . . . . .	III-102
Using Commands. . . . .	III-103
Usage Considerations. . . . .	III-103
Examples . . . . .	III-105
Computation. . . . .	III-122
Algorithms . . . . .	III-122
References. . . . .	III-122

## **3 Missing Value Analysis** **III-123**

Statistical Background. . . . .	III-123
Techniques for Handling Missing Values . . . . .	III-125
Randomness and Missing Data. . . . .	III-131
Testing for Randomness . . . . .	III-133
A Final Caution. . . . .	III-134
Missing Value Analysis in SYSTAT . . . . .	III-134
Missing Value Analysis Dialog Box . . . . .	III-134
Using Commands. . . . .	III-136
Usage Considerations. . . . .	III-137
Examples . . . . .	III-137
Computation. . . . .	III-183
Algorithms . . . . .	III-183
References. . . . .	III-184



## **4 Multidimensional Scaling**

**III-185**

Statistical Background . . . . .	III-186
Assumptions. . . . .	III-186
Collecting Dissimilarity Data . . . . .	III-187
Scaling Dissimilarities . . . . .	III-188
Multidimensional Scaling in SYSTAT . . . . .	III-189
Multidimensional Scaling Dialog Box . . . . .	III-189
Using Commands . . . . .	III-194
Usage Considerations . . . . .	III-194
Examples. . . . .	III-195
Computation . . . . .	III-210
Algorithms . . . . .	III-211
Missing Data . . . . .	III-212
References . . . . .	III-213

## **5 Multinormal Tests**

**III-215**

Statistical Background . . . . .	III-215
Multinormal Tests in SYSTAT . . . . .	III-216
Multinormal Tests Dialog Box . . . . .	III-216
Using Commands . . . . .	III-217
Usage Considerations . . . . .	III-217
Examples. . . . .	III-218
References . . . . .	III-221

## **6 Multivariate Analysis of Variance**

**III-223**

Statistical Background . . . . .	III-224
MANOVA Tests . . . . .	III-225
MANOVA in SYSTAT . . . . .	III-227

MANOVA: Estimate Model Dialog Box . . . . .	III-227
Hypothesis Test Dialog Box . . . . .	III-232
Between-Groups Testing . . . . .	III-239
Within-Group Testing . . . . .	III-241
Post hoc Test for Repeated measures. . . . .	III-242
Using Commands. . . . .	III-244
Usage Considerations. . . . .	III-244
Examples . . . . .	III-246
References . . . . .	III-259

## **7 Nonlinear Models**

**III-261**

Statistical Background. . . . .	III-262
Modeling the Dose-Response Function . . . . .	III-262
Loss Functions . . . . .	III-265
Model Estimation. . . . .	III-269
Problems . . . . .	III-269
Nonlinear Models in SYSTAT . . . . .	III-270
Nonlinear Regression: Estimate Model . . . . .	III-270
Loss Functions for Analytic Function Minimization. . . . .	III-281
Using Commands. . . . .	III-283
Usage Considerations. . . . .	III-283
Examples . . . . .	III-284
Computation. . . . .	III-316
Algorithms . . . . .	III-316
Missing Data . . . . .	III-316
References . . . . .	III-318

## **8 Nonparametric Tests**

**III-319**

Statistical Background. . . . .	III-320
---------------------------------	---------

Rank (Ordinal) Data . . . . .	III-320
Categorical (Nominal) Data . . . . .	III-321
Robustness . . . . .	III-321
Nonparametric Tests for Independent Samples in SYSTAT . . . . .	III-322
Kruskal-Wallis Test Dialog Box . . . . .	III-322
Two-Sample Kolmogorov-Smirnov Test Dialog Box . . . . .	III-323
Using Commands . . . . .	III-325
Nonparametric Tests for Related Variables in SYSTAT . . . . .	III-325
Sign Test Dialog Box . . . . .	III-325
Wilcoxon Signed-Rank Test Dialog Box . . . . .	III-326
Friedman Test Dialog Box . . . . .	III-328
Quade Test Dialog Box . . . . .	III-329
Using Commands . . . . .	III-331
Nonparametric Tests for Single Samples in SYSTAT . . . . .	III-331
One-Sample Kolmogorov-Smirnov Test Dialog Box . . . . .	III-331
Anderson-Darling Test Dialog Box . . . . .	III-334
Wald-Wolfowitz Runs Test Dialog Box . . . . .	III-337
Using Commands . . . . .	III-338
Usage Considerations . . . . .	III-339
Examples . . . . .	III-340
Computation . . . . .	III-355
Algorithms . . . . .	III-355
References . . . . .	III-355

## ***9 Partial Least Squares Regression*** ***III-357***

Statistical Background . . . . .	III-357
Model Building . . . . .	III-358
Cross-Validation . . . . .	III-360
Partial Least Squares Regression in SYSTAT . . . . .	III-361
Partial Least Squares Regression Dialog Box . . . . .	III-361
Using Commands . . . . .	III-364
Usage Considerations . . . . .	III-364

Examples . . . . .	III-365
Computation . . . . .	III-377
Algorithms . . . . .	III-377
Missing Data . . . . .	III-378
References . . . . .	III-378

## **10 Partially Ordered Scalogram Analysis with Coordinates** **III-381**

Statistical Background. . . . .	III-381
Coordinates . . . . .	III-383
POSAC in SYSTAT. . . . .	III-384
POSAC Dialog Box . . . . .	III-384
Using Commands. . . . .	III-385
Usage Considerations. . . . .	III-385
Examples . . . . .	III-386
Computation. . . . .	III-395
Algorithms . . . . .	III-395
Missing Data . . . . .	III-395
References . . . . .	III-395

## **11 Path Analysis (RAMONA)** **III-397**

Statistical Background. . . . .	III-397
The Path Diagram. . . . .	III-397
Path Analysis in SYSTAT. . . . .	III-405
Instructions for using RAMONA. . . . .	III-405
The MODEL statement. . . . .	III-407
RAMONA Options . . . . .	III-411
Usage Considerations. . . . .	III-413
Examples . . . . .	III-414



Computation . . . . .	III-452
RAMONA's Model . . . . .	III-452
Algorithms . . . . .	III-454
References . . . . .	III-460
Acknowledgments . . . . .	III-461

## ***Statistics IV***

### ***1 Perceptual Mapping IV-1***

Statistical Background . . . . .	IV-1
Preference Mapping. . . . .	IV-2
Biplots and MDPREF. . . . .	IV-6
Procrustes Rotations . . . . .	IV-7
Perceptual Mapping in SYSTAT . . . . .	IV-7
Perceptual Mapping Dialog Box . . . . .	IV-7
Using Commands . . . . .	IV-9
Usage Considerations . . . . .	IV-9
Examples. . . . .	IV-9
Computation . . . . .	IV-16
Algorithms . . . . .	IV-16
Missing data. . . . .	IV-16
References . . . . .	IV-16

### ***2 Power Analysis IV-19***

Statistical Background . . . . .	IV-20
Error Types . . . . .	IV-21
Power . . . . .	IV-22

Displaying Power Results . . . . .	IV-32
Generic Power Analysis . . . . .	IV-34
Power Analysis in SYSTAT. . . . .	IV-39
Single Proportion . . . . .	IV-39
Equality of Two Proportions . . . . .	IV-40
Single Correlation Coefficient . . . . .	IV-42
Equality of Two Correlation Coefficients . . . . .	IV-44
One-Sample z-test . . . . .	IV-46
Two-Sample z-test . . . . .	IV-48
One-Sample t-test. . . . .	IV-50
Paired t-test . . . . .	IV-51
Two-Sample t-test . . . . .	IV-53
One-Way ANOVA . . . . .	IV-55
Two-Way ANOVA. . . . .	IV-57
Generic Power Analysis . . . . .	IV-60
Using Commands. . . . .	IV-62
Usage Considerations. . . . .	IV-62
Examples . . . . .	IV-63
Computation. . . . .	IV-83
Algorithms . . . . .	IV-83
References. . . . .	IV-83

### **3 Probability Calculator**

### **IV-85**

Statistical Background. . . . .	IV-85
Probability Calculator in SYSTAT . . . . .	IV-86
Univariate Discrete Distributions Dialog Box . . . . .	IV-86
Univariate Continuous Distributions Dialog Box . . . . .	IV-87
Using Commands. . . . .	IV-90
Usage Considerations. . . . .	IV-90
Examples . . . . .	IV-90
References. . . . .	IV-98

## **4 Probit Analysis**

**IV-99**

Statistical Background . . . . .	IV-99
Interpreting the Results . . . . .	IV-100
Probit Analysis in SYSTAT . . . . .	IV-100
Probit Regression Dialog Box . . . . .	IV-100
Using Commands . . . . .	IV-103
Usage Considerations . . . . .	IV-103
Examples . . . . .	IV-104
Computation . . . . .	IV-107
Algorithms . . . . .	IV-107
Missing Data . . . . .	IV-107
References . . . . .	IV-107

## **5 Quality Analysis**

**IV-109**

Statistical Background . . . . .	IV-109
Quality Analysis in SYSTAT . . . . .	IV-110
Histogram . . . . .	IV-110
Quality Analysis: Histogram Dialog Box . . . . .	IV-110
Pareto Charts . . . . .	IV-111
Pareto Chart Dialog Box . . . . .	IV-112
Box-and-Whisker Plots . . . . .	IV-112
Box-and-Whisker Plot Dialog Box . . . . .	IV-113
Control Charts . . . . .	IV-114
Run Charts . . . . .	IV-114
Run Chart Dialog Box . . . . .	IV-115
Shewhart Control Charts . . . . .	IV-116
Shewhart Control Chart Dialog Box . . . . .	IV-116
OC and ARL curves . . . . .	IV-134
Operating Characteristic Curves . . . . .	IV-135
Operating Characteristic Curve Dialog Box . . . . .	IV-135
Average Run Length Curves . . . . .	IV-136

Average Run Length Dialog Box . . . . .	IV-137
Cusum Charts . . . . .	IV-142
Cumulative Sum Chart Dialog Box . . . . .	IV-142
Moving Average Charts . . . . .	IV-144
Moving Average Chart Dialog Box . . . . .	IV-144
Exponentially Weighted Moving Average Charts . . . . .	IV-146
Exponentially Weighted Moving Average Chart Dialog Box . . . . .	IV-146
X-MR Charts . . . . .	IV-149
X-MR Chart Dialog Box . . . . .	IV-150
Regression Charts . . . . .	IV-152
Regression Chart Dialog Box . . . . .	IV-152
TSQ Charts . . . . .	IV-153
TSQ Chart Dialog Box . . . . .	IV-154
Process Capability Analysis . . . . .	IV-155
Process Capability Analysis Dialog Box . . . . .	IV-159
Using Commands . . . . .	IV-161
Usage Considerations . . . . .	IV-162
Examples . . . . .	IV-163
References . . . . .	IV-217

## **6 Random Sampling**

**IV-219**

Statistical Background . . . . .	IV-220
Random Sampling in SYSTAT . . . . .	IV-220
Univariate Discrete Distributions Dialog Box . . . . .	IV-220
Univariate Continuous Distributions Dialog Box . . . . .	IV-222
Using Commands . . . . .	IV-223
Distribution Notations used in Random Sampling . . . . .	IV-223
Usage Considerations . . . . .	IV-224
Examples . . . . .	IV-225
Computation . . . . .	IV-228
Algorithms . . . . .	IV-228
References . . . . .	IV-228



## **7 Response Surface Methods** *IV-231*

Statistical Background . . . . .	IV-231
Fitting a Response Surface . . . . .	IV-232
Contour and Surface plot . . . . .	IV-233
Response Optimization . . . . .	IV-234
Response Surface Methods in SYSTAT. . . . .	IV-237
Response Surface Methods: Optimize Dialog Box . . . . .	IV-240
Using Commands . . . . .	IV-244
Usage Considerations . . . . .	IV-244
Examples. . . . .	IV-245
Computation . . . . .	IV-252
References . . . . .	IV-253

## **8 Robust Regression** *IV-255*

Statistical Background . . . . .	IV-256
Least Absolute Deviations (LAD) Regression . . . . .	IV-260
M Regression . . . . .	IV-261
Least Median Squares (LMS) Regression . . . . .	IV-261
Least Trimmed Squares (LTS) Regression . . . . .	IV-261
Scale (S) Regression . . . . .	IV-262
Rank Regression . . . . .	IV-262
Asymptotic Standard Errors, Confidence Intervals and Robust R2	IV-262
Robust Regression in SYSTAT . . . . .	IV-263
Least Absolute Deviation (LAD) Regression Dialog Box . .	IV-263
M Regression Dialog Box. . . . .	IV-265
Least Median of Squares (LMS) Regression Dialog Box . .	IV-268
Least Trimmed Squares (LTS) Regression Dialog Box . . .	IV-271
S Regression Dialog Box . . . . .	IV-275
Rank Regression Dialog Box . . . . .	IV-278
Using Commands . . . . .	IV-279
Usage Considerations . . . . .	IV-279

Examples . . . . .	IV-280
Computation . . . . .	IV-287
Algorithms . . . . .	IV-287
Missing Data . . . . .	IV-288
References . . . . .	IV-288

## ***9 Set and Canonical Correlations*** ***IV-291***

Statistical Background. . . . .	IV-291
Sets . . . . .	IV-292
Partialing . . . . .	IV-292
Notation. . . . .	IV-293
Measures of Association Between Sets. . . . .	IV-293
$R^2_{Y,X}$ Proportion of Generalized Variance . . . . .	IV-293
$T^2_{Y,X}$ and $P^2_{Y,X}$ Proportions of Additive Variance . . . . .	IV-294
Interpretations. . . . .	IV-295
Types of Association between Sets. . . . .	IV-296
Testing the Null Hypothesis . . . . .	IV-297
Estimates of the Population $R^2_{Y,X}$ , $T^2_{Y,X}$ , and $P^2_{Y,X}$ . . . . .	IV-299
Set and Canonical Correlations in SYSTAT . . . . .	IV-299
Set and Canonical Correlations Dialog Box . . . . .	IV-299
Category . . . . .	IV-301
Options . . . . .	IV-303
Using Commands. . . . .	IV-304
Usage Considerations. . . . .	IV-304
Examples . . . . .	IV-305
Computation. . . . .	IV-315
Algorithms . . . . .	IV-315
Missing Data . . . . .	IV-316
References . . . . .	IV-316

## **10 Signal Detection Analysis**

**IV-319**

Statistical Background . . . . .	IV-319
Detection Parameters . . . . .	IV-320
Signal Detection Analysis in SYSTAT . . . . .	IV-321
Signal Detection Analysis Dialog Box . . . . .	IV-321
Using Commands . . . . .	IV-324
Usage Considerations . . . . .	IV-325
Examples . . . . .	IV-328
Computation . . . . .	IV-346
Algorithms . . . . .	IV-346
Missing Data . . . . .	IV-346
References . . . . .	IV-346

## **11 Smoothing**

**IV-349**

Statistical Background . . . . .	IV-350
The Three Ingredients of Nonparametric Smoothers . . . . .	IV-350
A Sample Data Set . . . . .	IV-351
Kernels . . . . .	IV-352
Bandwidth . . . . .	IV-355
Smoothing Functions . . . . .	IV-358
Smoothness . . . . .	IV-360
Interpolation and Extrapolation . . . . .	IV-360
Close Relatives (Roses by Other Names) . . . . .	IV-360
Smoothing in SYSTAT . . . . .	IV-362
Smooth & Plot Dialog Box . . . . .	IV-362
Using Commands . . . . .	IV-366
Usage Considerations . . . . .	IV-366
Examples . . . . .	IV-367
References . . . . .	IV-382

## ***12 Spatial Statistics***

***IV-385***

Statistical Background. . . . .	IV-385
The Basic Spatial Model . . . . .	IV-385
The Geostatistical Model . . . . .	IV-387
Variogram. . . . .	IV-388
Variogram Models . . . . .	IV-389
Anisotropy . . . . .	IV-392
Simple Kriging . . . . .	IV-393
Ordinary Kriging . . . . .	IV-394
Universal Kriging. . . . .	IV-394
Simulation . . . . .	IV-394
Point Processes . . . . .	IV-395
Spatial Statistics in SYSTAT . . . . .	IV-399
Spatial Statistics Dialog Box . . . . .	IV-399
Using Commands. . . . .	IV-408
Usage Considerations. . . . .	IV-410
Examples . . . . .	IV-411
Computation. . . . .	IV-426
Missing Data . . . . .	IV-426
Algorithms . . . . .	IV-426
References . . . . .	IV-426

## ***13 Survival Analysis***

***IV-427***

Statistical Background. . . . .	IV-428
Graphics . . . . .	IV-429
Parametric Modeling . . . . .	IV-432
Survival Analysis in SYSTAT . . . . .	IV-435
Survival Analysis: Nonparametric Dialog Box. . . . .	IV-436
Survival Analysis: Parametric and Cox Dialog Box . . . . .	IV-439
Using Commands. . . . .	IV-447



Usage Considerations . . . . .	IV-448
Examples. . . . .	IV-449
Computation . . . . .	IV-476
Algorithms . . . . .	IV-476
Missing Data . . . . .	IV-476
References . . . . .	IV-484

## ***14 Test Item Analysis***

***IV-487***

Statistical Background . . . . .	IV-488
Classical Model . . . . .	IV-489
Latent Trait Model . . . . .	IV-490
Test Item Analysis in SYSTAT . . . . .	IV-491
Classical Test Item Analysis Dialog Box . . . . .	IV-491
Logistic Test Item Analysis Dialog Box . . . . .	IV-493
Using Commands . . . . .	IV-494
Usage Considerations . . . . .	IV-495
Examples. . . . .	IV-498
Computation . . . . .	IV-506
Algorithms . . . . .	IV-506
Missing Data . . . . .	IV-507
References . . . . .	IV-507

## ***15 Time Series***

***IV-509***

Statistical Background . . . . .	IV-510
Smoothing. . . . .	IV-510
ARIMA Modeling and Forecasting. . . . .	IV-514
Seasonal Decomposition and Adjustment . . . . .	IV-523
Exponential Smoothing . . . . .	IV-524
Trend Analysis . . . . .	IV-525

Fourier Analysis . . . . .	IV-526
Graphical Displays for Time Series in SYSTAT . . . . .	IV-528
Time Axis Format Dialog Box . . . . .	IV-528
Time Series Plot Dialog Box . . . . .	IV-529
ACF Plot Dialog Box . . . . .	IV-529
PACF Plot Dialog Box . . . . .	IV-530
CCF Plot Dialog Box . . . . .	IV-531
Using Commands . . . . .	IV-532
Transformations of Time Series in SYSTAT . . . . .	IV-532
Transform Dialog Box . . . . .	IV-532
Clear Series . . . . .	IV-534
Using Commands . . . . .	IV-534
Smoothing a Time Series in SYSTAT . . . . .	IV-535
Moving Average Smoothing Dialog Box . . . . .	IV-535
LOWESS Smoothing Dialog Box . . . . .	IV-536
Exponential Smoothing Dialog Box . . . . .	IV-537
Using Commands . . . . .	IV-539
Seasonal Adjustments in SYSTAT . . . . .	IV-539
Seasonal Adjustment Dialog Box . . . . .	IV-539
Using Commands . . . . .	IV-540
ARIMA Models in SYSTAT . . . . .	IV-540
ARIMA Dialog Box . . . . .	IV-540
Using Commands . . . . .	IV-542
Trend Analysis in SYSTAT . . . . .	IV-542
Trend Analysis dialog box . . . . .	IV-542
Using Commands . . . . .	IV-544
Fourier Models in SYSTAT . . . . .	IV-544
Fourier Transformation Dialog Box . . . . .	IV-545
Using Commands . . . . .	IV-546
Usage Considerations . . . . .	IV-546
Examples . . . . .	IV-547
Computation . . . . .	IV-578
Algorithms . . . . .	IV-578
References . . . . .	IV-578

## **16 Two-Stage Least Squares** **IV-581**

Statistical Background . . . . .	IV-581
Two-Stage Least Squares Estimation . . . . .	IV-582
Heteroskedasticity . . . . .	IV-583
Two-Stage Least Squares in SYSTAT . . . . .	IV-584
Two-Stage Least Squares Regression Dialog Box . . . . .	IV-584
Using Commands . . . . .	IV-586
Usage Considerations . . . . .	IV-586
Examples . . . . .	IV-587
Computation . . . . .	IV-597
Algorithms . . . . .	IV-597
Missing Data . . . . .	IV-597
References . . . . .	IV-597

## **Acronym & Abbreviation Expansions**

## **Index**

---

# *List of Examples*

Multi Way: Standardize Tables . . . . .	I-291
A Model with Interaction . . . . .	II-315
A Nested-Factorial Model with Case Frequencies . . . . .	II-412
Actuarial Life Tables . . . . .	IV-453
Additive Trees . . . . .	I-120
AIC and Schwarz's BIC . . . . .	III-258
Analysis of Covariance (ANCOVA) . . . . .	II-209
Analysis of Covariance . . . . .	II-153
Anderson-Darling Test . . . . .	III-353
ANOVA Assumptions and Contrasts . . . . .	II-126
ARIMA Models . . . . .	IV-566
ARL Curve . . . . .	IV-197
Autocorrelation Plot . . . . .	IV-548
Automatic Stepwise Regression . . . . .	II-71
Basic Statistics for Rows . . . . .	I-340
Basic Statistics . . . . .	I-324
Bayesian Regression . . . . .	II-99



Binary Logit with Interactions . . . . .	III-33
Binary Logit with Multiple Predictors . . . . .	III-27
Binary Logit with One Predictor . . . . .	III-24
Binary Profiles . . . . .	III-388
Bonferroni and Dunn-Sidak adjustments . . . . .	I-552
Box-and-Whisker Plots . . . . .	IV-166
Box-Behnken Design . . . . .	I-380
Box-Cox Model . . . . .	I-143
Box-Hunter Fractional Factorial Design . . . . .	I-373
By-Choice Data Format . . . . .	III-69
c Chart . . . . .	IV-191
Calculating Percentiles Using Inverse Cumulative Distribution Function . . . . .	IV-93
Calculating Probability Mass Function and Cumulative Distribution Function for Discrete Distributions . . . . .	IV-90
Canonical Correlation Analysis . . . . .	II-246
Canonical Correlations: Using Text Output . . . . .	I-33
Canonical Correlations—Simple Model . . . . .	IV-305
Casewise Pattern Table . . . . .	III-142
Categorical Variables and Clustered Data . . . . .	II-449
Central Composite Response Surface Design . . . . .	I-384
Chi-Square Model for Signal Detection . . . . .	IV-340

Choice Data . . . . .	I-136
Circle Model . . . . .	IV-11
Classical Test Analysis . . . . .	IV-498
Classification Tree . . . . .	I-55
Clustered Data in Mixed Regression . . . . .	II-442
Cochran's Test of Linear Trend . . . . .	I-273
Comparing Correlation Estimation Methods . . . . .	III-168
Computation of p-value Using 1-CF Function . . . . .	IV-94
Conditional Logistic Regression. . . . .	III-56
Confidence Curves and Regions. . . . .	III-287
Confidence Interval for Non-Centrality Parameter in One-Way Balanced Fixed Effect ANOVA . . . . .	IV-95
Confidence Intervals for Mean and Median . . . . .	I-28
Confidence Intervals for One-Way Table Percentages . . . . .	I-250
Confidence Intervals for Smoothers . . . . .	IV-368
Confidence Intervals. . . . .	II-414
Contingency Table Analysis. . . . .	IV-312
Contouring the Loss Function . . . . .	III-296
Contrasts . . . . .	I-435
Correlation Estimation. . . . .	III-154

Correspondence Analysis (Simple) . . . . .	I-207
Covariance Alternatives to Repeated Measures . . . . .	II-234
Cox Regression . . . . .	IV-462
Cross-Correlation Plot . . . . .	IV-550
Crossover and Changeover Designs . . . . .	II-222
Cross-Validation . . . . .	I-444
Cross-Validation . . . . .	III-371
Cumulative Histogram . . . . .	IV-164
Cusum Charts . . . . .	IV-201
Deciles of Risk and Model Diagnostics . . . . .	III-39
Density Clustering Examples . . . . .	I-112
Differencing . . . . .	IV-552
Discrete Choice Models . . . . .	III-60
Discriminant Analysis Using Automatic Backward Stepping . . . . .	I-420
Discriminant Analysis Using Automatic Forward Stepping . . . . .	I-413
Discriminant Analysis Using Complete Estimation . . . . .	I-409
Discriminant Analysis Using Interactive Stepping . . . . .	I-427
Discriminant Analysis . . . . .	II-238
Employment Discrimination . . . . .	I-147
Equality of Proportions . . . . .	IV-63

Estimation: ML and REML . . . . .	II-369
EWMA Chart . . . . .	IV-204
Exploring with Residuals . . . . .	II-334
Factor Analysis Using a Covariance Matrix. . . . .	I-482
Factor Analysis Using a Rectangular File . . . . .	I-485
Fine Tuning . . . . .	II-382
Fisher's Exact Test. . . . .	I-271
Fitting a Second Order Response Surface . . . . .	IV-245
Fitting Binomial Distribution . . . . .	I-504
Fitting Discrete Uniform Distribution . . . . .	I-505
Fitting Exponential Distribution . . . . .	I-507
Fitting Gumbel Distribution . . . . .	I-508
Fitting Multiple Distributions . . . . .	I-513
Fitting Normal Distribution . . . . .	I-510
Fitting Weibull Distribution . . . . .	I-511
Fixing Parameters and Evaluating Fit . . . . .	III-290
Flexible Beta Linkage Method for Hierarchical Clustering . . . . .	I-115
Fourier Modeling of Temperature . . . . .	IV-575
Fractional Factorial Design . . . . .	I-372
Fractional Factorial Designs . . . . .	II-213



Frequency Input . . . . .	I-256
Friedman Test for the Case with Ties . . . . .	III-348
Friedman Test . . . . .	III-347
From VC to MIXED . . . . .	II-357
Full Factorial Designs . . . . .	I-371
Functions of Parameters . . . . .	III-293
Gamma Model for Signal Detection . . . . .	IV-344
Geometric Mean . . . . .	I-326
Getting Acquainted with the Output Layout . . . . .	II-311
Guttman Loss Function. . . . .	III-198
Hadi Robust Outlier Detection . . . . .	I-192
Harmonic Mean. . . . .	I-327
Heteroskedasticity-Consistent Standard Errors . . . . .	IV-587
Hierarchical Clustering with Leaf Option . . . . .	I-118
Hierarchical Clustering: Clustering Cases . . . . .	I-105
Hierarchical Clustering: Clustering Variables and Cases . . . . .	I-109
Hierarchical Clustering: Clustering Variables . . . . .	I-108
Hierarchical Clustering: Distance Matrix Input . . . . .	I-111
Histogram. . . . .	IV-163
Hotelling's T-Square . . . . .	II-237

Hypothesis testing . . . . .	II-372
Hypothesis Testing . . . . .	III-77
Incomplete Block Designs. . . . .	II-212
Independent Samples t-Test . . . . .	IV-72
Individual Differences Multidimensional Scaling. . . . .	III-200
Interactive Stepwise Regression . . . . .	II-75
Internal Model . . . . .	IV-12
Iterated Principal Axis. . . . .	I-476
Iteratively Reweighted Least-Squares for Logistic Models . . . . .	III-299
Kinetic Models. . . . .	III-313
K-Means Clustering . . . . .	I-96
Kriging (Ordinary). . . . .	IV-411
Kruskal Method . . . . .	III-195
Kruskal-Wallis Test . . . . .	III-340
Latin Square Designs . . . . .	II-220
Latin Squares . . . . .	I-375
Least-Squares Regression . . . . .	I-23
Life Tables: The Kaplan-Meier Estimator. . . . .	IV-449
Logistic Model (One Parameter) . . . . .	IV-500
Logistic Model (Two Parameter) . . . . .	IV-503

Logistic Model for Signal Detection . . . . .	IV-335
Loglinear Modeling of a Four-Way Table . . . . .	III-105
Longitudinal Data in Mixed Regression . . . . .	II-457
LOWESS Smoothing . . . . .	IV-558
Mann-Kendall test . . . . .	IV-572
Mann-Whitney Test . . . . .	III-342
Mantel-Haenszel Test . . . . .	I-293
Maximum Likelihood Estimation . . . . .	III-298
Maximum Likelihood . . . . .	I-473
McNemar's Test of Symmetry . . . . .	I-277
Minimizing an Analytic Function . . . . .	III-315
Missing Category Codes . . . . .	I-257
Missing Cells Designs (the Means Model) . . . . .	II-224
Missing Data . . . . .	II-340
Missing Data: EM Estimation . . . . .	I-186
Missing Data: Pairwise Deletion . . . . .	I-185
Missing Value Imputation . . . . .	III-176
Missing Values: Preliminary Examinations . . . . .	III-137
Mixture Design with Constraints . . . . .	I-382
Mixture Design . . . . .	I-381

Mixture Models . . . . .	II-247
Moving Average Chart . . . . .	IV-203
Moving Averages . . . . .	IV-555
Multinomial Logit . . . . .	III-50
Multiple Categories . . . . .	III-390
Multiple Correspondence Analysis . . . . .	I-214
Multiple Linear Regression . . . . .	II-67
Multiple Response Optimization using Desirability Analysis. . . . .	IV-250
Multiplicative Seasonal Factor . . . . .	IV-560
Multiplicative Seasonality with a Linear Trend . . . . .	IV-561
Multivariate Layout for Longitudinal Data . . . . .	II-473
Multivariate Nested Design . . . . .	III-253
Multivariate Normality Assessment of Anthropometric Measurements . . . . .	III-219
Multivariate Normality Assessment of Perspiration Measurements . . . . .	III-218
Multivariate Regression by PLS Technique. . . . .	III-368
Multiway Tables. . . . .	I-279
Negative Exponential Model for Signal Detection . . . . .	IV-336
Nested Designs . . . . .	II-215
Nested Effects . . . . .	II-320
Nested Random Effects . . . . .	II-417



Nesting in Design Structure . . . . .	II-402
Nesting in treatment structure . . . . .	II-399
Nesting versus Crossing . . . . .	II-408
Nonlinear Model with Three Parameters . . . . .	III-284
Nonmetric Unfolding . . . . .	III-203
Nonparametric Model for Signal Detection . . . . .	IV-333
Nonparametric: One Sample Kolmogorov-Smirnov Test Statistic. . . . .	I-36
Normal Distribution Model for Signal Detection . . . . .	IV-328
Normality Assessment Using Shapiro-Wilk and Anderson-Darling Test . . . . .	I-341
np Chart. . . . .	IV-183
N-tiles and P-tiles. . . . .	I-338
OC Curve for Binomial Distribution . . . . .	IV-199
OC Curve for Variances . . . . .	IV-198
OC Curve . . . . .	IV-197
Odds Ratios. . . . .	I-269
One-Sample Kolmogorov-Smirnov Test for Non-Central Chi-square Distribution . . . . .	III-352
One-Sample Kolmogorov-Smirnov Test for Normal Distribution . . . . .	III-350
One-Sample t-Test . . . . .	I-547
One-Sample z-Test . . . . .	I-544
One-Way ANOVA and Sample Size Estimation. . . . .	IV-77

One-Way ANOVA . . . . .	II-122
One-Way ANOVA . . . . .	II-203
One-Way MANOVA . . . . .	III-246
One-Way Repeated Measures . . . . .	II-155
One-Way Tables . . . . .	I-248
Optimal Designs: Coordinate Exchange. . . . .	I-386
Optimizing Response using Canonical Analysis . . . . .	IV-247
Optimum Choice of Number of Factors . . . . .	III-375
Outliers in X-space and Y-space . . . . .	IV-284
Outliers in X-space . . . . .	IV-283
Outliers in Y-space . . . . .	IV-280
p Chart . . . . .	IV-189
Paired t-Test . . . . .	I-548
Paired t-Test . . . . .	IV-67
Pairwise comparisons . . . . .	II-145
Pareto Charts. . . . .	IV-165
Partial Autocorrelation Plot . . . . .	IV-549
Partial Correlations . . . . .	II-248
Partial Set Correlation Model . . . . .	IV-308
Path Analysis and Standard Errors . . . . .	III-442

Path Analysis Basics . . . . .	III-414
Path Analysis Using Rectangular Input . . . . .	III-434
Path Analysis with a Restart File . . . . .	III-419
PCA with Beta Distribution . . . . .	IV-215
PCA With Box-Cox Transformation . . . . .	IV-213
PCA with Normal Distribution . . . . .	IV-212
PDL with Instrumental Variables . . . . .	IV-596
PDL without Instrumental Variables . . . . .	IV-595
Pearson Correlations . . . . .	I-179
Percentages . . . . .	I-258
Piecewise Regression . . . . .	III-311
Plackett-Burman Design . . . . .	I-379
Point Statistics . . . . .	IV-418
Poisson Model for Signal Detection . . . . .	IV-342
Poisson Test . . . . .	I-551
Polynomial Regression and Smoothing . . . . .	IV-370
POSAC: Proportion of Profile Pairs Correctly Represented . . . . .	I-34
Post hoc tests . . . . .	II-379
Power Scaling Ratio Data . . . . .	III-208
Prediction of New Observations . . . . .	II-95

Principal Components Analysis (Within Groups) . . . . .	II-242
Principal Components . . . . .	I-469
Probabilities Associated with Correlations . . . . .	I-188
Probit Analysis (Simple Model) . . . . .	IV-104
Probit Analysis with Interactions . . . . .	IV-106
Procrustes Rotation . . . . .	IV-14
Quade Test for Cases with Ties . . . . .	III-349
Quade Test for Multiple Comparisons. . . . .	III-349
Quadratic Model. . . . .	I-438
Quantiles. . . . .	III-45
R Chart. . . . .	IV-180
Randomized Block Designs . . . . .	II-211
Regression Charts . . . . .	IV-207
Regression Imputation. . . . .	III-181
Regression Tree with Box Plots . . . . .	I-57
Regression Tree with Dit Plots . . . . .	I-59
Regression using SSCP, Covariance or Correlation matrices. . . . .	II-89
Regression with Ecological or Grouped Data . . . . .	II-86
Regression without the Constant . . . . .	II-87
Regression . . . . .	III-306



Repeated Measures Analysis in the Presence of Subject-Specific Covariates	III-255
Repeated Measures Analysis of Covariance . . . . .	II-170
Repeated Measures ANOVA for One Grouping Factor and One Within Factor with Ordered Levels. . . . .	II-160
Repeated Measures ANOVA for Two Grouping Factors and One Within Factor . . . . .	II-163
Repeated Measures ANOVA for Two Trial Factors . . . . .	II-166
Repeated Measures Experiment with Covariates. . . . .	II-366
Residuals and Diagnostics for Simple Linear Regression . . . . .	II-63
Ridge Analysis . . . . .	IV-249
Ridge Regression Analysis . . . . .	II-97
Robust Discriminant Analysis . . . . .	I-449
Robust Estimation (Measures of Location) . . . . .	III-301
Rotation. . . . .	I-478
Run Chart. . . . .	IV-167
s chart. . . . .	IV-178
S2 and S3 Coefficients . . . . .	I-196
Sampling Distribution of Double Exponential (Laplace) Median . . . . .	IV-225
Saving Basic Statistics: Multiple Statistics and Grouping Variables . . . . .	I-328
Saving Basic Statistics: One Statistic and One Grouping Variable . . . . .	I-327
Scalogram Analysis—A Perfect Fit . . . . .	III-386

Screening Effects . . . . .	III-114
Seasonal Trend tests . . . . .	IV-573
Seemingly Unrelated Regression Equations. . . . .	II-91
Separate Variance Hypothesis Tests. . . . .	II-151
Sign and Wilcoxon Tests for Multiple Variables . . . . .	III-346
Sign Test . . . . .	III-343
Simple Correspondence Analysis using Raw Data . . . . .	I-212
Simple Linear Regression . . . . .	II-55
Simulation of Assembly System. . . . .	IV-226
Simulation . . . . .	IV-417
Single-Degree-of-Freedom Designs . . . . .	II-148
Smart Correspondence Analysis with Row-by-Column Data . . . . .	I-210
Smoothing (4253H Filter) . . . . .	IV-557
Smoothing Binary Data in Three Dimensions. . . . .	IV-380
Smoothing: Saving and Plotting Results . . . . .	IV-367
Spearman Correlations. . . . .	I-195
Spearman Rank Correlation . . . . .	I-27
Split Plot Design . . . . .	II-323
Split Plot Designs . . . . .	II-217
Stem-and-Leaf Plot for Rows . . . . .	I-342

Stem-and-Leaf Plot . . . . .	I-333
Stepwise Regression . . . . .	III-70
Stepwise Regression . . . . .	IV-468
Stratified Cox Regression . . . . .	IV-464
Stratified Kaplan-Meier Estimation . . . . .	IV-455
Structural Zeros. . . . .	III-117
Structured Covariance Matrix for Random Errors . . . . .	II-362
Tables with Ordered Categories . . . . .	I-275
Tables without Analyses . . . . .	III-121
Tackling different data format in Logistic Regression . . . . .	III-81
Taguchi Design . . . . .	I-377
Test for Equality of Several Variances . . . . .	I-558
Test for Equality of Two Correlation Coefficients . . . . .	I-562
Test for Equality of Two Proportions . . . . .	I-564
Test for Equality of Two Variances . . . . .	I-557
Test for Single Proportion . . . . .	I-564
Test for Single Variance . . . . .	I-556
Test for Specific Correlation Coefficient. . . . .	I-560
Test for Zero Correlation Coefficient . . . . .	I-559
Testing Nonzero Null Hypotheses . . . . .	II-85

Testing whether a Single Coefficient Equals Zero . . . . .	II-81
Testing whether Multiple Coefficients Equal Zero . . . . .	II-83
Tetrachoric Correlation . . . . .	I-198
The Nelson-Aalen Estimator . . . . .	IV-451
The Weibull Model for Fully Parametric Analysis . . . . .	IV-472
Time Series Plot . . . . .	IV-547
Transformations . . . . .	I-182
Transformations . . . . .	II-60
Treatment or design? . . . . .	II-406
TSLS without lag and with hypothesis testing . . . . .	IV-593
TSQ Chart . . . . .	IV-209
Turnbull Estimation: K-M for Interval-Censored Data . . . . .	IV-459
Two-Sample t-Test . . . . .	I-549
Two-Sample z-Test . . . . .	I-545
Two-Stage Instrumental Variables . . . . .	IV-592
Two-Stage Least Squares . . . . .	IV-590
Two-Way MANOVA . . . . .	III-248
Two-Way ANOVA . . . . .	II-132
Two-way ANOVA . . . . .	IV-80
Two-Way Table Measures (Long Results) . . . . .	I-263



Two-Way Table Measures . . . . .	I-261
Two-Way Tables . . . . .	I-253
u Chart . . . . .	IV-195
Unbalanced ANOVA . . . . .	II-146
Unbalanced Data: Different Types of ANOVA . . . . .	II-328
Univariate Regression by PLS Technique . . . . .	III-365
Unordered Data . . . . .	I-198
Unusual Distances . . . . .	IV-424
Usefulness of Jackknife estimate . . . . .	I-30
Using Covariates . . . . .	II-326
Validity indices RMSSTD, Pseudo F, and Pseudo T-square with cities . . . . .	I-116
Variance Chart . . . . .	IV-176
Vector Model . . . . .	IV-9
Wald-Wolfowitz Runs Test . . . . .	III-354
Weighting Means . . . . .	II-234
Wilcoxon Test . . . . .	III-345
Within-Group Testing . . . . .	III-257
Word Frequency . . . . .	I-140
X-bar Chart . . . . .	IV-168
X-MR Chart (Sigma Estimation with Median) . . . . .	IV-206

# Logistic Regression

## Our Statistics and Quality Tools

(Compiled by Nandini Jayaraman and Joseph V. Pech)

Logit module estimates parameters for logistic, multinomial, conditional, and discrete choice models. For each model you fit, Logit reports the parameter estimates, confidence interval for the parameters,  $\chi^2$  test, Wald tests, standard errors for odds ratio, confidence interval for the odds ratio, and coefficient of determination of parameter estimates. Logit performs Wald test, likelihood ratio test, backward and forward stepwise regression. Logit also provides graphical diagnostic, prediction success and classification matrix, average and marginal derivatives, model-based simulation of response surface, derivative and score statistics to specify start values and to separate data into training and test samples, robust standard errors, control of significance levels for coefficient testing, dependent variable recoding, dependent variable coding, choice of reference category, variable selection, variable evaluation, and integrated plotting facility. Coefficient estimates of the model are stored in a file. Available Information Criterion (AIC) and Bayesian Information Criterion (BIC) are used for model selection. For more information on AIC, see Parzen, "Generalized Linear Models," "Variable Selection" on page 13 in *Statistical Theory in Planning and Science* (1992).

Logit module also provides the following features:

- Estimation of the parameters of the logistic regression model using the maximum likelihood method.
- Estimation of the parameters of the multinomial logistic regression model using the maximum likelihood method.
- Estimation of the parameters of the conditional logistic regression model using the maximum likelihood method.
- Estimation of the parameters of the discrete choice model using the maximum likelihood method.
- Calculation of the odds ratio and its confidence interval.
- Calculation of the Wald test, likelihood ratio test, and score test.
- Calculation of the standard errors for the parameter estimates.
- Calculation of the coefficient of determination.
- Calculation of the average and marginal derivatives.
- Calculation of the prediction success and classification matrix.
- Calculation of the model-based simulation of response surface.
- Calculation of the derivative and score statistics.
- Calculation of the robust standard errors.
- Calculation of the control of significance levels for coefficient testing.
- Calculation of the dependent variable recoding.
- Calculation of the dependent variable coding.
- Calculation of the choice of reference category.
- Calculation of the variable selection.
- Calculation of the variable evaluation.
- Calculation of the integrated plotting facility.
- Calculation of the coefficient estimates of the model.
- Calculation of the AIC and BIC for model selection.

Many of the results generated by the Logit module can be saved to data files for future use. The results can be saved in a file in case of binary logistic regression. The results can be saved in a file in case of multinomial logistic regression. The results can be saved in a file in case of conditional logistic regression. The results can be saved in a file in case of discrete choice model.



# Logistic Regression

Dan Steinberg and Phillip Colla

(revised by Nandita Ingawale and Avijit Maji)

LOGIT module estimates parameters for binary, multinomial, conditional, and discrete choice models. For each model you fit, LOGIT reports the parameter estimates, confidence interval for the parameters, z ratio, odds ratio, standard errors for odds ratio, confidence interval for the odds ratio, and correlation matrix of parameter estimates. Logit performs Wald test, score tests, forward, backward and interactive stepwise regression. Logit also produces Pregibon regression diagnostics, prediction success and classification tables, independent variable derivatives, model-based simulation of response curves, deciles of risk tables, options to specify start values and to separate data into learning and test samples, robust standard errors, control of significance levels for confidence interval calculations, zero/one dependent variable coding, choice of reference group in automatic dummy variable generation, and integrated plotting tools. Output includes information criteria values (Akaike Information Criterion (AIC) and Schwarz's BIC) which are tools for model selection. For more information on AIC and BIC see Chapter 1: Linear Models, "Variable Selection" on page 15 in *Statistics I* and Burnham and Anderson (1992).

Many of the results generated by modeling, testing, or diagnostic procedures can be saved to data files for subsequent graphing and display with the graphics routines. In case of binary logistic regression, SYSTAT displays the area under the curve and receiver operating characteristic (ROC) curve as Quick Graph.



## ***Statistical Background***

The LOGIT module is SYSTAT's comprehensive program for logistic regression analysis and provides tools for model building, model evaluation, prediction, simulation, hypothesis testing, and regression diagnostics. The program is designed to be easy for the novice and can produce the results most analysts need with just three simple commands. In addition, many advanced features are also included for sophisticated research projects. Beginners can skip over any unfamiliar concepts and gradually increase their mastery of logistic regression by working through the tools incorporated here.

LOGIT will estimate binary (Cox and Snell, 1989), multinomial (Anderson, 1972), conditional logistic regression models (Breslow and Day, 1980), and the discrete choice model (Luce, 2005; McFadden, 1973). The LOGIT framework is designed for analyzing the determinants of a categorical dependent variable. Typically, the dependent variable is binary and coded as 0 or 1; however, it may be multinomial and coded as an integer ranging from 1 to  $k$  or 0 to  $k - 1$ .

Studies you can conduct with LOGIT include bioassay, epidemiology of disease (cohort or case-control), clinical trials, market research, transportation research (mode of travel), psychometric studies, and voter-choice analysis. The LOGIT module can also be used to analyze ranked choice information once the data have been suitably transformed (Beggs, Cardell, and Hausman, 1981).

This chapter contains a brief introduction to logistic regression and a description of the commands and features of the module. If you are unfamiliar with logistic regression, the textbook by Hosmer and Lemeshow (2000) is an excellent place to begin; Breslow and Day (1980) provide an introduction in the context of case-control studies; Train (1986) and Ben-Akiva and Lerman (1985) introduce the discrete-choice model for econometrics; Wrigley (2002) discusses the model for geographers; and Hoffman and Duncan (1988) review discrete choice in a demographic-sociological context. Valuable surveys appear in Amemiya (1981), McFadden (1976, 1982, 1984), and Maddala (1986).

### ***Binary Logit***

Although logistic regression may be applied to any categorical dependent variable, it is most frequently seen in the analysis of binary data, in which the dependent variable takes on only two values. Examples include survival beyond five years in a clinical

trial, presence or absence of disease, responding to a specified dose of a toxin, voting for a political candidate, and participating in the labor force.

In modeling the conditional distribution of the response variable  $Y$ , given the independent variable(s)  $X$ , we choose an appropriate characteristic of the conditional distribution which depends on the independent variables in an explicable manner. Thus in linear regression it is the expected value, in survival analysis it is the hazard rate and in logit (or probit) analysis it is  $Prob(Y=1 | x)$ .

When  $Y$  and  $X$  are positively associated,  $Prob(Y=1 | x)$  is an increasing function of  $x$ , it lies between 0 and 1 and so the obviously appropriate model is a distribution function  $F(x)$ . In logit analysis, the logistic distribution function is used to model  $Prob(Y=1 | x)$ . Now with  $m$  and  $s$  as the location and scale parameters respectively, the distribution function is

$$F(x) = F_0\left(\frac{x-\mu}{\sigma}\right)$$

where  $F_0$  is the standard logistic distribution function given by

$$F_0(x) = \frac{\exp(x)}{1 + \exp(x)}$$

It is convenient to write

$$\begin{aligned} F(x) &= F_0(\alpha + \beta x) \\ &= \frac{\exp(\alpha + \beta x)}{1 + \exp(\alpha + \beta x)} \end{aligned}$$

$$\text{with } \alpha = -\frac{\mu}{\sigma} \text{ and } \beta = \frac{1}{\sigma}.$$

With more than one independent variable and not necessarily with positive association among them, the model in its general form is written as:

$$Prob(Y = 1 | \underline{x}) = \frac{\exp(\beta_0 + \underline{\beta}'\underline{x})}{1 + \exp(\beta_0 + \underline{\beta}'\underline{x})}$$

where an underline denotes the vector form. It can be easily seen that

$$\log \frac{Prob(Y = 1 | \underline{x})}{1 - Prob(Y = 1 | \underline{x})} = \beta_0 + \underline{\beta}'\underline{x}$$

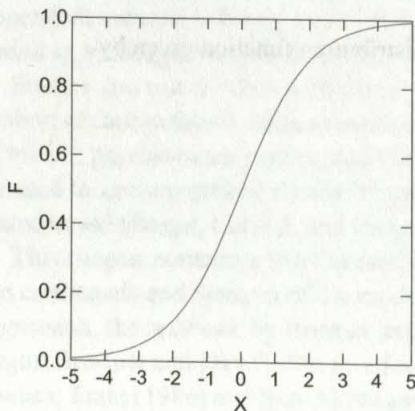
For data  $\{(y_i, x_i), i=1,2,3 \dots n\}$ , SYSTAT finds estimates of the parameters  $\beta_0$  and  $\beta$  using the maximum likelihood method of estimation.

In probit analysis the function  $F(x)$  is the cumulative distribution function of the normal distribution with  $m$  and  $s$  as the location and scale parameters respectively.

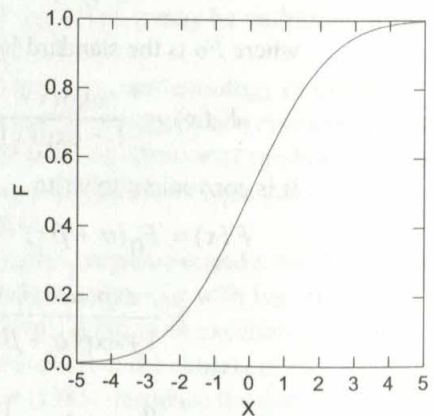
Logit analysis and probit analysis are quite similar in nature, the two curves also are alike with some difference in the shape, the logistic distribution having somewhat heavier tails. Whether to choose logit or probit will mostly depend on the nature of the phenomenon which gives rise to the data under consideration.

You can visually make the comparison from the following two graphs:

Logistic Distribution Function



Normal Distribution Function



You may notice that while plotting the normal distribution function, we have taken the standard deviation  $s = 1.81$  which is also the standard deviation of standard logistic distribution.



## **Multinomial Logit**

**Multinomial logit** is a logistic regression model having a dependent variable with more than two levels (Agresti, 2002; Santer and Duffy, 2004; Nerlove and Press, 1973). Examples of such dependent variables include political preference (Democrat, Republican, Independent), health status (healthy, moderately impaired, seriously impaired), smoking status (current smoker, former smoker, never smoked), and job classification (executive, manager, technical staff, clerical, other). Outside of the difference in the number of levels of the dependent variable, the multinomial logit is very similar to the binary logit, and most of the standard tools of interpretation, analysis, and model selection can be applied. In fact, the polytomous unordered logit we discuss here is essentially a combination of several binary logits estimated simultaneously (Begg and Gray, 1984). We use the term **polytomous** to differentiate this model from the conditional logistic regression and discrete choice models discussed below.

There are important differences between binary and multinomial models. Chiefly, the multinomial output is more complicated than that of the binary model, and care must be taken in the interpretation of the results. Fortunately, LOGIT provides some new tools that make the task of interpretation much easier. There is also a difference in dependent variable coding. The binary logit dependent variable is normally coded 0 or 1, whereas the multinomial dependent can be coded 1, 2, ...,  $k$ , (that is, it starts at 1 rather than 0) or 0, 1, 2, ...,  $k - 1$ .

## **Conditional Logit**

The conditional logistic regression model has become a major analytical tool in epidemiology since the work of Prentice and Breslow (1978), Breslow et al. (1978), Prentice and Pyke (1979), and the extended treatment of case-control studies in Breslow and Day (1980). A mathematically similar model with the same name was introduced independently and from a rather different perspective by McFadden (1973) in econometrics. The models have since seen widespread use in the considerably different contexts of biomedical research and social science, with parallel literatures on sampling, estimation techniques, and statistical results. In epidemiology, conditional logit is used to estimate relative risks in matched sample case-control studies (Breslow, 1982), whereas in econometrics a similar likelihood function is used to model consumer choices as a function of the attributes of alternatives. We begin this section with a treatment of the biomedical use of the conditional logistic model. A separate



section on the discrete choice model covers the econometric version and contains certain fine points that may be of interest to all readers. A discussion of parallels in the two literatures appears in Steinberg (1991).

In the traditional conditional logistic regression model, you are trying to measure the risk of disease corresponding to different levels of exposure to risk factors. The data have been collected in the form of matched sets of cases and controls, where the cases have the disease, the controls do not, and the sets are matched on background variables such as age, sex, marital status, education, residential location, and possibly other health indicators. The matching variables combine to form strata over which relative risks are to be estimated; thus, for example, a small group of persons of a given age, marital status, and health history will form a single stratum. The matching variables can also be thought of as proxies for a larger set of unobserved background variables that are assumed to be constant within strata. The logit for the  $j$ th individual in the  $i$ th stratum can be written as:

$$\text{logit}(p_{ij}) = a_i + b_j X_{ij}$$

where  $X_{ij}$  is the vector of exposure variables and  $a_i$  is a parameter dedicated to the stratum. Since case-control studies will frequently have a large number of small matched sets, the  $a_i$  are nuisance parameters that can cause problems in estimation (Cox and Hinkley, 1979). In the example discussed below, there are 63 matched sets, each consisting of one case and four controls, with information on seven exposure variables for every subject.

The problem with estimating an unconditional model for these data is that we would need to include  $63 - 1 = 62$  dummy variables for the strata. This would leave us with possibly 70 parameters being estimated for a data set with only 315 observations. Furthermore, increasing the sample size will not help because an additional stratum parameter would have to be estimated for each additional matched set in the study sample. By working with the appropriate conditional likelihood, however, the nuisance parameters can be eliminated, simplifying estimation and protecting against potential biases that may arise in the unconditional model (Cox, 1975; Chamberlain, 1980). The conditional model requires estimation only of the relative risk parameters of interest.

LOGIT allows the estimation of models for matched sample case-control studies with one case and any number of controls per set. Thus, matched pair studies, as well as studies with varying numbers of controls per case, are easily handled. However, not all commands discussed so far are available for conditional logistic regression.

## Discrete Choice Logit

Econometricians and psychometricians have developed a version of logit frequently called the **discrete choice model**, or **McFadden's conditional logit model** (McFadden, 1973, 1976, 1982, 1984; Hensher and Johnson, 1981; Ben-Akiva and Lerman, 1985; Train, 1986; Luce, 2005). This multinomial model differs from the standard polytomous logit in the interpretation of the coefficients, the number of parameters estimated, the syntax of the model sentence, and options for data layout.

The discrete choice framework is designed specifically to model an individual's choices in response to the characteristics of the choices. Characteristics of choices are attributes such as price, travel time, horsepower, or calories; they are features of the alternatives that an individual might choose from. By contrast, characteristics of the chooser, such as age, education, income, and marital status, are attributes of a person.

The classic application of the discrete choice model has been to the choice of travel mode to work (Domencich and McFadden, 1975). Suppose a person has three alternatives: private auto, car pool, and commuter train. The individual is assumed to have a utility function representing the desirability of each option, with the utility of an alternative depending solely on its own characteristics. With travel time and travel cost as key characteristics determining mode choice, the utility of each option could be written as:

$$U_i = B_1 T_i + B_2 C_i + e_i$$

where  $i = 1, 2, 3$  represents private auto, car pool, and train, respectively. In this random utility model, the utility  $U_i$  of the  $i$ th alternative is determined by the travel time  $T_i$ , the cost  $C_i$  of that alternative, and a random error term,  $e_i$ . Utility of an alternative is assumed not to be influenced by the travel times or costs of other alternatives available, although choice will be determined by the attributes of all available alternatives. In addition to the alternative characteristics, utility is sometimes also determined by an alternative specific constant.

The choice model specifies that an individual will choose the alternative with the highest utility as determined by the equation above. Because of the random component, we are reduced to making statements concerning the probability that a given choice is made. If the error terms are distributed as i.i.d. extreme value, it can be shown that the probability of the  $i$ th alternative being chosen is given by the familiar logit formula.



$$Prob(U_i > U_j \text{ for all } j \neq i) = \frac{\exp(X_i b)}{\sum \exp(X_i b)}$$

Suppose that for the first few cases our data are as follows:

Subject	Choice	Auto(1)	Auto(2)	Pool(1)	Pool(2)	Train(1)	Train(2)	Sex	Age
1	1	20	3.50	35	2.00	65	1.10	Male	27
2	3	45	6.00	65	3.00	65	1.00	Female	35
3	1	15	1.00	30	0.50	60	1.00	Male	22
4	2	60	5.50	70	2.00	90	2.00	Male	45
5	3	30	4.25	40	1.75	55	1.50	Male	52

The third record has a person who chooses to go to work by private auto (choice = 1); when he drives, it takes 15 minutes to get to work and costs one dollar. Had he carpooled instead, it would have taken 30 minutes to get to work and cost 50 cents. The train would have taken an hour and cost one dollar. For this case, the utility of each option is given by

$$\begin{aligned} U_{(\text{private auto})} &= b_1 * 15 + b_2 * 1.00 + \text{error}_{13} \\ U_{(\text{car pool})} &= b_1 * 30 + b_2 * 0.50 + \text{error}_{23} \\ U_{(\text{train})} &= b_1 * 60 + b_2 * 1.00 + \text{error}_{33} \end{aligned}$$

The error term has two subscripts, one pertaining to the alternative and the other pertaining to the individual. The error is individual-specific and is assumed to be independent of any other error or variable in the data set. The parameters  $b_1$  and  $b_2$  are common utility weights applicable to all individuals in the sample. In this example, these are the only parameters, and their number does not depend on the number of alternatives individuals can choose from. If a person also had the option of walking to work, we would expand the model to include this alternative with

$$U_{(\text{walking})} = b_1 * 70 + b_2 * 0.00 + \text{error}_{43}$$

and we would still be dealing with only the two regression coefficients  $b_1$  and  $b_2$ .

This highlights a major difference between the discrete choice and standard polytomous logit models. In polytomous logit, the number of parameters grows with the number alternatives; if the value of NCAT (number of categories) is increased from 3 to 4, a whole new vector of parameters is estimated. By contrast, in the discrete choice model without a constant, increasing the number of alternatives does not increase the number of discrete choice parameters estimated.

Finally, we need to look at the *optional* constant. Optional is emphasized because it is perfectly legitimate to estimate without a constant, and, in certain circumstances, it is even necessary to do so. If we were to add a constant to the travel mode model, we would obtain the following utility equations:

$$U_i = b_{oi} + b_1 T_i + b_2 C_i + e_i$$

where  $i = 1, 2, 3$  represents private auto, car pool, and train, respectively. The constant here,  $b_{oi}$ , is alternative-specific, with a separate one estimated for each alternative:  $b_{o1}$  corresponds to private auto;  $b_{o2}$ , to car pooling; and  $b_{o3}$ , to train. Like polytomous logit, the constant pertaining to the reference group is normalized to 0 and is not estimated.

An alternative specific CONSTANT is entered into a discrete choice model to capture unmeasured desirability of an alternative. Thus, the first constant could reflect the convenience and comfort of having your own car (or in some cities the inconvenience of having to find a parking space), and the second might reflect the inflexibility of schedule associated with shared vehicles. With NCAT=3, the third constant will be normalized to 0.

### ***Stepwise Logit***

Automatic model selection can be extremely useful for analyzing data with a large number of covariates for which there is little or no guidance from previous research. For these situations, LOGIT supports stepwise regression, allowing forward, backward, mixed, and interactive covariate selection, with full control over forcing, selection criteria, and candidate variables (including interactions). The procedure is based on Peduzzi, Holford, and Hardy (1980).

Stepwise regression results in a model that cannot be readily evaluated using conventional significance criteria in hypothesis tests, but the model may prove useful for prediction. We strongly suggest that you separate the sample into learning and test sets for assessment of predictive accuracy before fitting a model to the full data set. See the cautionary discussion and references in Statistics II, Chapter 2 Linear Models I: Linear Regression.

## Logistic Regression in SYSTAT

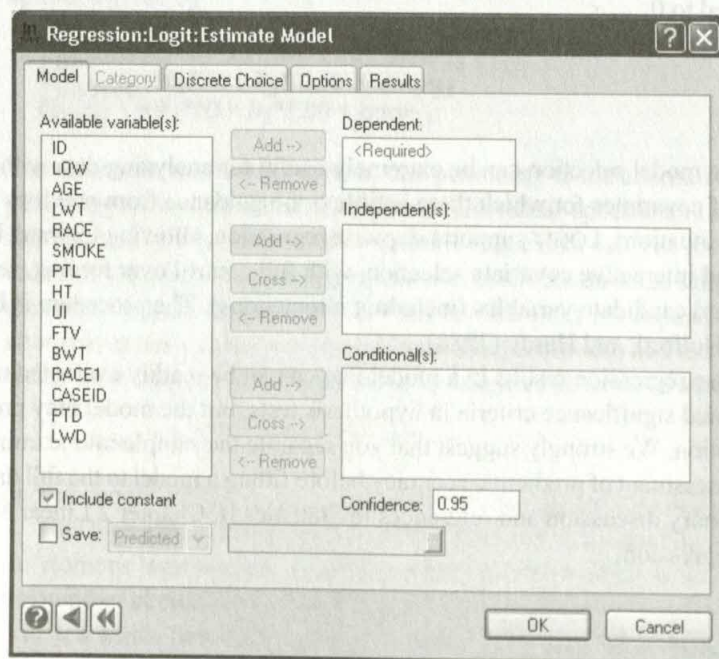
### Estimate Model Dialog Box

Logistic regression analysis provides tools for model building, model evaluation, prediction, simulation, hypothesis testing, and regression diagnostics.

Many of the results generated by modeling, testing, or diagnostic procedures can be saved to SYSTAT data files for subsequent graphing and display. New data handling features for the discrete choice model allow tremendous savings in disk space when choice attributes are constant, and in some models, performance is greatly improved.

To open the Logit Regression: Estimate Model dialog box, from the menus choose:

Analyze  
Regression  
Logit  
Estimate Model...





**Dependent.** Select the variable you want to examine. The dependent variable should be a categorical numeric variable.

**Independent(s).** Select one or more continuous or categorical variables. To add an interaction to your model, use the Cross button. For example, to add the term *SMOKE\*LWT*, add *SMOKE* to the Independent list and then add *LWT* by clicking Cross.

**Conditional(s).** Select conditional variables. To add interactive conditional variables to your model, use the Cross button. For example, to add the term *SMOKE\*LWT*, add *SMOKE* to the Conditional list and then add *LWT* by clicking Cross.

**Include constant.** The constant is an optional parameter. Deselect Include constant check box to obtain a model through the origin. When in doubt, include the constant.

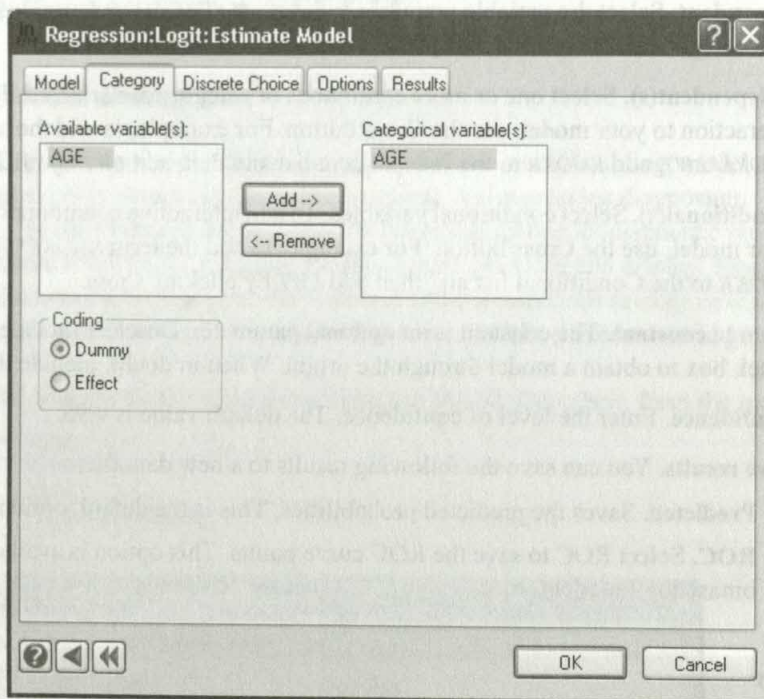
**Confidence.** Enter the level of confidence. The default value is 0.95.

**Save results.** You can save the following results to a new data file:

- **Predicted.** Saves the predicted probabilities. This is the default option.
- **ROC.** Select ROC to save the ROC curve points. This option is available only for binary logit models.

### Category

You must specify numeric or string grouping variables that define cells. Specify for all categorical variables for which logistic regression analysis should generate design variables.



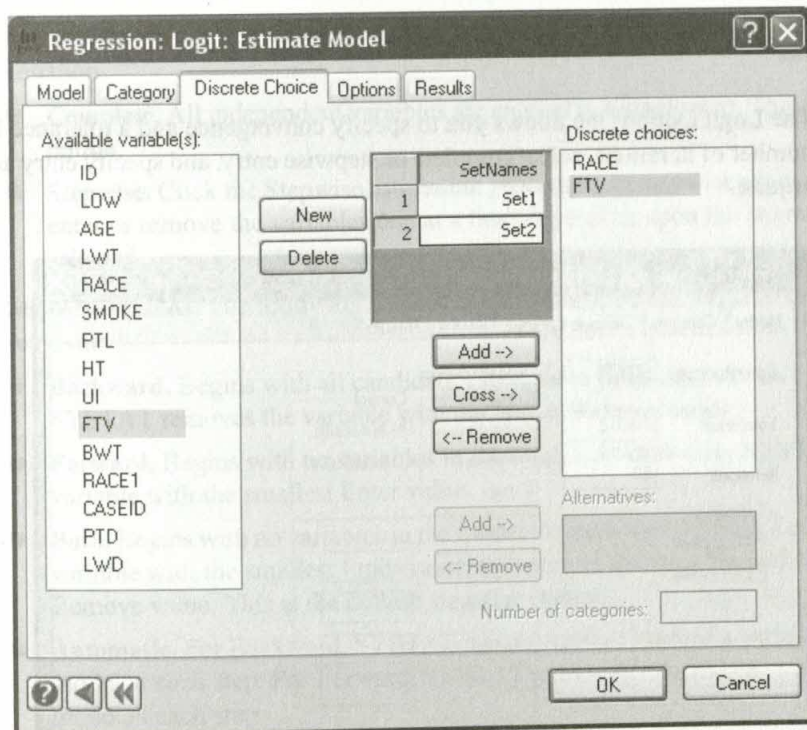
**Categorical variable(s).** Categorize an independent variable when it has several categories; for example, education levels, which could be divided into the following categories: less than high school, some high school, finished high school, some college, finished bachelor's degree, finished master's degree, and finished doctorate. On the other hand, a variable such as age in years would not be categorical unless age were broken up into categories such as under 21, 21–65, and over 65.

**Coding.** You must indicate the coding method to apply to categorical variables. The two available options include:

- **Dummy.** Produces dummy codes for the design variables instead of effect codes. Coding of dummy variables is the classic analysis of variance parameterization, in which the sum of effects estimated for a classifying variable is 0. If your categorical variable has  $k$  categories,  $k - 1$  dummy variables are created. This is the default coding option.
- **Effect.** Click Effect to produce parameter estimates that are differences from group means.

### Discrete Choice

The discrete choice framework is designed specifically to model an individual's choices in response to the characteristics of the choices. Characteristics of choices are attributes such as price, travel time, horsepower, or calories; they are features of the alternatives that an individual might choose from. You can define set names for groups of variables, and create, edit, or delete variables.



**SetNames.** Specifies conditional variables. Enter a set name and then you can add and cross variables. To create a new set, click New. Repeat this process until you have defined all of your sets. You can edit existing sets by highlighting the name of the set in the SetNames drop-down list. To delete a set, select the set in the drop-down list and click Delete. When you click OK, SYSTAT will check that each set name has a definition. If a set name exists but no variables were assigned to it, the set is discarded and the set name will not be in the drop-down list when you return to this dialog box.



**Alternatives.** Specify an alternative for discrete choice. Characteristics of choice are features of the alternatives that an individual might choose between. It is needed only when the number of alternatives in a choice model varies per subject.

**Number of categories.** Specify the number of categories or alternatives the variable has. This is needed only for the by-choice data layout where the values of the dependent variable are not explicitly coded. This is only enabled when the Alternatives field is not empty.

### Options

The Logit Options tab allows you to specify convergence and a tolerance level, and number of iterations, select complete or stepwise entry, and specify entry and removal criteria.

The screenshot shows the 'Regression:Logit:Estimate Model' dialog box with the 'Options' tab selected. The dialog has five tabs: 'Model', 'Category', 'Discrete Choice', 'Options', and 'Results'. The 'Options' tab contains the following settings:

- Convergence:**
- Tolerance:**
- Iterations:**
- Estimation:**
  - ☒ Complete
  - ☐ Stepwise
- Stepwise options:**
  - Direction:**
    - ☐ Backward
    - ☐ Forward
    - ☒ Both
  - Control:**
    - ☒ Automatic
    - ☐ Interactive
- Probability:**
  - Enter:**
  - Remove:**
  - Maximum steps:**
  - Force:**

At the bottom of the dialog, there are navigation buttons (a question mark, a left arrow, and a right arrow) and 'OK' and 'Cancel' buttons.

**Convergence.** Enter the largest relative change in any coordinate before iterations terminate.

**Tolerance.** Enter a value that prevents the entry of a variable that is highly correlated with the independent variables already included in the model. Enter a value between 0 and 1. Typical values are 0.01 or 0.001. The higher the value (closer to 1), the lower the correlation required to exclude a variable.

**Iterations.** Enter the maximum number of iterations for fitting your model.

**Estimation.** To control the method used to enter and remove variables from the equation.

- **Complete.** All independent variables are entered in a single step. This is the default option.
- **Stepwise.** Click the Stepwise estimation procedure. In stepwise procedure you can enter or remove the variables one at a time depending upon the stepwise options selected.

**Stepwise options.** The following alternatives are available for stepwise entry and removal:

- **Backward.** Begins with all candidate variables in the model. At each step, SYSTAT removes the variable with the largest Remove value.
- **Forward.** Begins with no variables in the model. At each step, SYSTAT adds the variable with the smallest Enter value.
- **Both.** Begins with no variables in the model. At each step, SYSTAT either adds the variable with the smallest Enter value, or removes the variable with the largest Remove value. This is the default stepwise option.
- **Automatic.** For Backward, SYSTAT automatically removes a variable from your model at each step. For Forward, SYSTAT automatically adds a variable to the model at each step.
- **Interactive.** At each step in the model building, you select the variable to enter into or remove from the model.

**Probability.** You can also control the criteria used to enter variables into and remove variables from the model:

- **Enter.** Enter the probability to enter variable(s) into the model. The variable is entered into the model if its alpha value is less than the specified value. Enter a value between 0 and 1 (for example, 0.025). The default value is 0.15.



- **Remove.** Enter the probability to remove variable(s) into the model. The variable is removed from the model if its alpha value is greater than the specified value. Enter a value between 0 and 1 (for example, 0.025). The default value is 0.15.

**Maximum steps.** Enter the maximum number of steps.

**Force.** Enter the number of variable. Forces the first  $n$  variables listed in your model to remain in the equation.

## Results

The screenshot shows a software dialog box titled "Regression: Logit: Estimate Model". It has five tabs: "Model", "Category", "Discrete Choice", "Options", and "Results". The "Results" tab is selected. Inside the dialog, there are several options and input fields:

- ☐ Robust standard errors
- ☐ Prediction success table
- ☐ Classification table
- ☐ Means
- ☐ Derivatives
  - ☒ Individual
  - ☐ Average
- ☐ Deciles of risk
  - ☒ Based on probability values: [0.1 0.2 0.3 0.4 0.5 0.6 0.7 0.8]
  - ☐ Based on equal counts per bin
- Number of bins: [ ]
- ☐ Save residuals: [ ]
- Cutoff: [0.5]

At the bottom, there are navigation buttons (back, forward, search) and "OK" and "Cancel" buttons.

**Robust standard errors:** Select the Robust standard errors check box for the robust standard error of parameter estimates when the model to be estimated by maximum likelihood is misspecified.

**Prediction success table.** Select the Prediction success table check box which summarizes the classificatory power of the model.

**Classification table:** Select the Classification table check box that summarizes the results of your fitted model based on a cutoff point.

- **Cutoff.** Enter the desired cutoff point for displaying the classification table at that cutoff point. The default value is 0.5. The edit box is enabled only for binary logit models.

**Means.** Select the Means check box. It displays the average value for the variables in the model.

**Derivatives.** Select Derivatives check box. You can select the following options to produce a derivative table:

- **Individual.** Evaluates the change in the probability of outcome in response to a change in the covariate values. This is the default option.
- **Average.** Click Average to evaluate derivatives at the sample average of the covariates.

**Deciles of Risk.** After you successfully estimate your model using logistic regression, you can calculate deciles of risk. This feature is available only for binary logit models. This will help you make sure that your model fits the data and that the results are not unduly influenced by a handful of unusual observations. In using the deciles of risk table, please note that the goodness-of-fit statistics will depend on the grouping rule specified.

Two grouping rules are available:

- **Based on probability values.** Probability is reallocated across the possible values of the dependent variable as the independent variable changes. It provides a global view of covariate effects that is not easily seen when considering each binary submodel separately. In fact, the overall effect of a covariate on the probability of an outcome can be of the opposite sign of its coefficient estimate in the corresponding submodel. This is because the submodel concerns only two of the outcomes, whereas the derivative table considers all outcomes at once. By default, SYSTAT considers probability values from 0.1 to 1 in increments of 0.1. You may change these values.
- **Based on equal counts per bin.** Allocates approximately equal numbers of observations to each cell. Enter the number of cells or bins in the Number of bins text box.

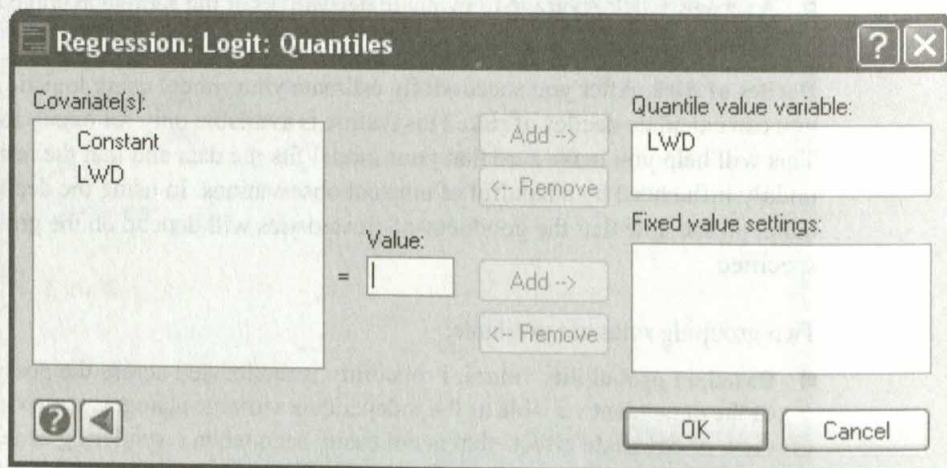
**Save residuals.** Saves the residuals to new data file.

## Quantiles

After estimating your model, you can calculate quantiles for any single-predictor in the model. This feature is available only for binary logit models. Quantiles of unadjusted data can be useful in assessing the suitability of a functional form when you are interested in the unconditional distribution of the failure times.

To open the Logit Regression: Quantiles dialog box, from the menus choose:

Analyze  
Regression  
Logit  
Quantiles...



**Covariate(s).** The Covariate(s) list contains all of the variables specified in the Independent list in the Model tab of Logit Regression: Estimate Model dialog box. You can set any of the covariates to a fixed value by selecting the variable in the Covariates list and entering a value in the Value text box. This constraint appears as variable name = value in the Fixed value settings list after you click Add. The quantiles for the desired variable correspond to a model in which the covariates are fixed at these values. Any covariates not fixed to a value are assigned the value of 0.

**Quantile value variable.** By default, the first variable in the Independent variable list in the Model tab of Logit Regression: Estimate Model dialog box is shown in this field.



You can change this to any variable from the list. This variable name is then issued as the argument for the QNTL command.

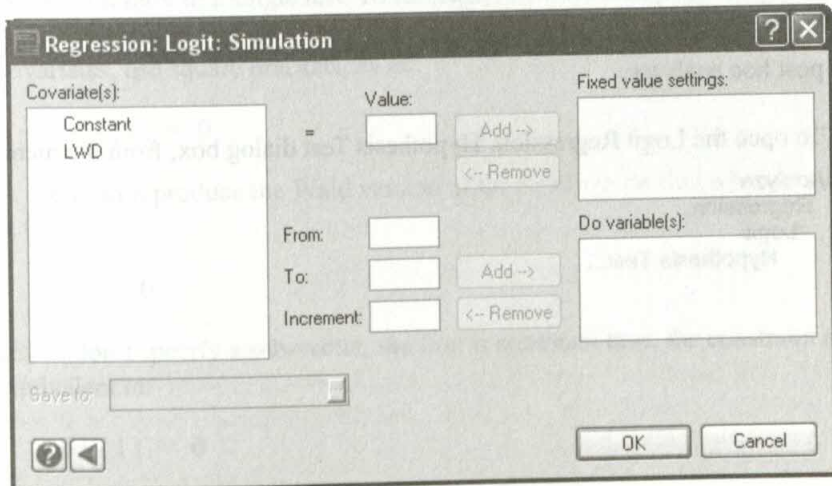
**Fixed value settings.** This box lists the fixed values on the covariates from which the logits are calculated.

## Simulation

SYSTAT allows you to generate and save predicted probabilities and odds ratios, using the last model estimated to evaluate a set of logits. The logits are calculated from a combination of fixed covariate values and a grid of values taken by some of the covariates as specified by you in the dialog box shown below.

To open the Logit Regression: Simulation dialog box, from the menus choose:

Analyze  
Regression  
Logit  
Simulation...



**Covariate(s).** The Covariate(s) list contains all of the variables specified in the Independent list on the Model tab of Logit Regression: Estimate Model dialog box. Select a covariate, enter a fixed value for the covariate in the Value text box, and click

the Add button corresponding to the Fixed value settings list. You can also specify a range of values for a covariate by entering the From, To and Increment values, and clicking the Add button corresponding to the Do variable(s) list.

**Value.** Enter the value at which the selected covariate should be fixed.

**Fixed value settings.** This box lists the fixed values on the covariates from which the logits are calculated.

**From.** Enter the starting value of the selected covariate.

**To.** Enter the ending value of the selected covariate.

**Increment.** Enter the increment for each step.

**Do variable(s).** This box lists the grid of values over which some or all of the covariates should vary.

When you specify a grid of values for one or more of the covariates, or when the model is multinomial, or when the dependent variable is a string variable, you should specify a file to which the simulation results will be saved.

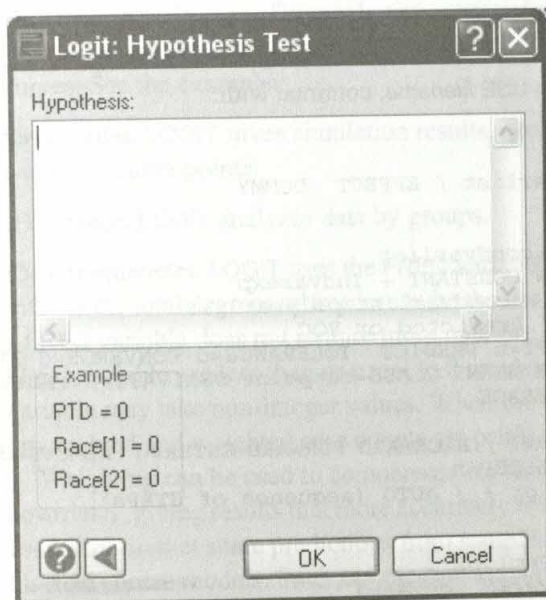
## Hypothesis

After you successfully estimate your model using logistic regression, you can perform post hoc analyses.

To open the Logit Regression: Hypothesis Test dialog box, from the menus choose:

- Analyze
- Regression
- Logit
- Hypothesis Test...





Enter the hypotheses that you would like to test. All the hypotheses that you list will be tested jointly in a single test. To test each restriction individually, you will have to revisit this dialog box each time. To reference dummies generated from categorical covariates, use square brackets, as in:

$$RACE[1] = 0$$

You can reproduce the Wald version of the z ratio by testing whether a coefficient is 0:

$$AGE = 0$$

If you don't specify a sub-vector, the first is assumed; thus, the constraint above is equivalent to:

$$AGE\{1\} = 0$$

24.2.2015  
14622



## Using Commands

After selecting a file with *USE filename*, continue with:

```

USE FILENAME
LOGIT
  CATEGORY grpvarlist / EFFECT DUMMY
  NCAT n
  ALT var
  SET parameter=condvarlist
  MODEL depvar = CONSTANT + indvarexp
        depvar = condvarlist; polyvarlist
  SAVE filename/ Predicted or ROC
  ESTIMATE /CONF $\bar{I}$ =u PREDICT TOLERANCE=d CONVERGE=d ITER=n
          RSE MEANS CLASS=cutpoint DERIVATIVE=INDIVIDUAL or
          AVERAGE
          or
          START / BACKWARD FORWARD ENTER=d REMOVE=d FORCE=n
          MAXSTEP=n
  STEP var or + or - / AUTO (sequence of STEPs)
  STOP
  SAVE
  DC / BINS=n P=p1,p2,...
  QNTL var / covar=d covar=d
  SIMULATE var1=d1, var2=d2, ... / DO var1=d1,d2,d3, var2=d1,d2,d3
  HYPOTHESIS
  CONSTRAIN argument
  TEST

```

## Usage Considerations

**Types of data.** LOGIT uses rectangular data only. The dependent variable is automatically taken to be categorical. To change the order of the categories, use the ORDER statement. For example,

```
ORDER CLASS / SORT=DESCENDING
```

LOGIT can also handle categorical predictor variables. Use the CATEGORY statement to create them, and use the EFFECTS or DUMMY options of CATEGORY to determine the coding method. Use the ORDER command to change the order of the categories.

**Print options.** For PLENGTH SHORT, the output gives N, the different strength of association, parameter estimates, confidence interval and associated tests. PLENGTH LONG gives, in addition to the above results, a correlation matrix of the parameter estimates.

**Quick Graphs.** In case of binary logistic regression, logit produces ROC curve as quick graph. Use the saved files from ESTIMATE or DC to produce diagnostic plots and fitted curves. See the examples.

**Saving files.** LOGIT saves simulation results, quantiles, or residuals, predicted values and ROC curve points.

**BY groups.** LOGIT analyzes data by groups.

**Case frequencies.** LOGIT uses the FREQ variable, if present, to weight cases. This inflates the total degrees of freedom to be the sum of the number of frequencies. Using a FREQ variable does not require more memory, however. Cases whose value on the FREQ variable are less than or equal to 0 are deleted from the analysis. The FREQ variable may take non-integer values. When the FREQ command is in effect, separate unweighted and weighted case counts are printed.

Weighting can be used to compensate for sampling schemes that stratify on the covariates, giving results that more accurately reflect the population. Weighting is also useful for market share predictions from samples stratified on the outcome variable in discrete choice models. Such samples are known as choice-based in the econometric literature (Manski and Lerman, 1977; Manski and McFadden, 1980; Coslett, 1980) and are common in matched-sample case-control studies where the cases are usually over-sampled, and in market research studies where persons who choose rare alternatives are sampled separately.

**Case weights.** LOGIT does not allow case weighting.



## Examples

The following examples begin with the simple binary logit model and proceed to more complex multinomial and discrete choice logit models. Along the way, we will examine diagnostics and other options used for applications in various fields.

### Example 1

#### Binary Logit with One Predictor

To illustrate the use of binary logistic regression, we take this example from Hosmer and Lemeshow's book *Applied Logistic Regression*, referred to below as H&L. Hosmer and Lemeshow (2000) consider data on low infant birth weight (*LOW*) as a function of several risk factors. These include the mother's age (*AGE*), mother's weight during last menstrual period (*LWT*), race (*RACE* = 1: white, *RACE* = 2: black, *RACE* = 3: other), smoking status during pregnancy (*SMOKE*), history of premature labor (*PTL*), hypertension (*HT*), uterine irritability (*UI*), and number of physician visits during first trimester (*FTV*). The dependent variable is coded 1 for birth weights less than 2500 grams and coded 0 otherwise. These variables have previously been identified as associated with low birth weight in the obstetrical literature.

The first model considered is the simple regression of *LOW* on a constant and *LWD*, a dummy variable coded 1 if *LWT* is less than 110 pounds and coded 0 otherwise. (See H&L, Table 3.17.) *LWD* and *LWT* are similar variable names. Be sure to note which is being used in the models that follow.

The input is:

```
USE HOSLEM
LOGIT
MODEL LOW=CONSTANT+LWD
ESTIMATE
```

The output is:

Logistic Regression

Categorical values encountered during processing are

Variables	Levels
LOW (2 levels)	0.000 1.000

Binary LOGIT Analysis

```
Dependent Variable : LOW
Input Records      : 189
```



Records for Analysis : 189

**Sample Split****Category Choices**

0 (REFERENCE)	130
1 (RESPONSE)	59
Total	189

**Log-Likelihood Iteration History**

Log-Likelihood at Iteration1	-131.005
Log-Likelihood at Iteration2	-113.231
Log-Likelihood at Iteration3	-113.121
Log-Likelihood at Iteration4	-113.121
Log-Likelihood	-113.121

**Information Criteria**

AIC	230.241
Schwarz's BIC	236.725

**Parameter Estimates**

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval Lower	Upper
1 CONSTANT	-1.054	0.188	-5.594	0.000	-1.423	-0.685
2 LWD	1.054	0.362	2.914	0.004	0.345	1.762

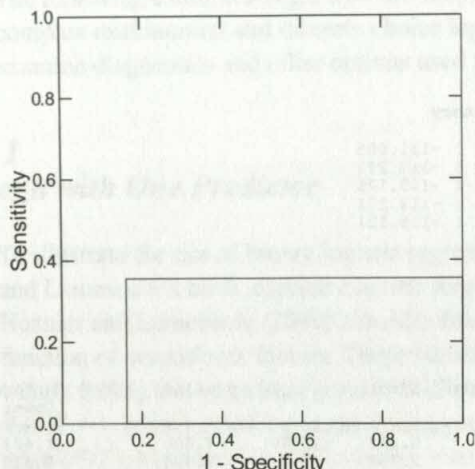
**Odds Ratio Estimates**

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval Lower	Upper
2 LWD	2.868	1.037	1.412	5.826

Log-Likelihood of Constants only Model = LL(0) : -117.336  
 2\*[LL(N)-LL(0)] : 8.431  
 df : 1  
 p-value : 0.004

McFadden's Rho-squared : 0.036  
 Cox and Snell R-square : 0.044  
 Naglekerke's R-square : 0.061

Receiver Operating Characteristic Curve



Area under ROC Curve : 0.597

The output begins with a listing of the dependent variable and the sample split between 0 (reference) and 1 (response) for the dependent variable. A brief iteration history follows, showing the progress of the procedure to convergence. Finally, the parameter estimates, standard errors, standardized coefficients (popularly called  $z$  ratios),  $p$  values, 95% confidence intervals, and ratios and the log-likelihood are presented.

### Coefficients

We can evaluate these results much like a linear regression. The coefficient on *LWD* is large relative to its standard error ( $z$  ratio = 2.914) and so appears to be an important predictor of low birth weight. The interpretation of the coefficient is quite different from ordinary regression, however. The logit coefficient tells how much the logit increases for a unit increase in the independent variable, but the probability of a 0 or 1 outcome is a nonlinear function of the logit.

## Odds Ratio

The odds-ratio table provides a more intuitively meaningful quantity for each coefficient. The odds of the response are given by  $p/(1-p)$ , where  $p$  is the probability of response, and the odds ratio is the multiplicative factor by which the odds change when the independent variable increases by one unit. In the first model, being a low-weight mother increases the odds of a low birth weight baby by a multiplicative factor of 2.868, with lower and upper confidence bounds of 1.41 and 5.83 and with standard error of odds ratio=1.037, respectively. Since the lower bound is greater than 1, the variable appears to represent a genuine risk factor. See Kleinbaum, Kupper, and Chambliss (1982) for a discussion.

## Example 2

### Binary Logit with Multiple Predictors

The binary logit example contains only a constant and a single dummy variable. We consider the addition of the continuous variable *AGE* to the model.

The input is:

```
USE HOSLEM
LOGIT
MODEL LOW=CONSTANT+LWD+AGE
ESTIMATE / MEANS
```

The output is:

#### Logistic Regression

Categorical values encountered during processing are

Variables	Levels
LOW (2 levels)	0.000 1.000

#### Binary LOGIT Analysis

```
Dependent Variable : LOW
Input Records      : 189
Records for Analysis : 189
```

#### Sample Split

Category Choices	
0 (REFERENCE)	130
1 (RESPONSE)	59
Total	189

## Independent Variable Means

PARAMETER	0	-1	OVERALL
1 CONSTANT	1.000	1.000	1.000
2 LWD	0.356	0.162	0.222
3 AGE	22.305	23.662	23.238

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-131.005
Log-Likelihood at Iteration2	-112.322
Log-Likelihood at Iteration3	-112.144
Log-Likelihood at Iteration4	-112.143
Log-Likelihood at Iteration5	-112.143
Log-Likelihood	-112.143

## Information Criteria

AIC	230.287
Schwarz's BIC	240.012

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	-0.027	0.762	-0.035	0.972	-1.521	1.467
2 LWD	1.010	0.364	2.773	0.006	0.296	1.724
3 AGE	-0.044	0.032	-1.373	0.170	-0.107	0.019

## Odds Ratio Estimates

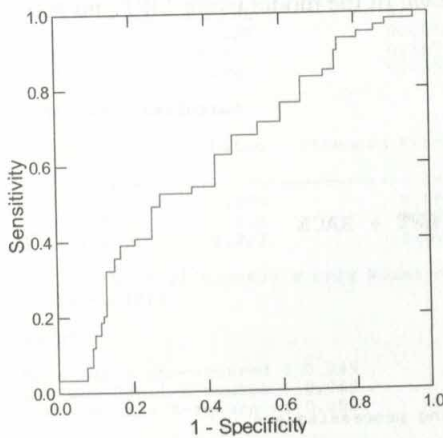
Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
2 LWD	2.746	1.000	1.345	5.607
3 AGE	0.957	0.031	0.898	1.019

Log-Likelihood of Constants only Model = LL(0) : -117.336  
 2\*[LL(N)-LL(0)] : 10.385  
 df : 2  
 p-value : 0.006

McFadden's Rho-squared : 0.044  
 Cox and Snell R-square : 0.053  
 Naglekerke's R-square : 0.075



Receiver Operating Characteristic Curve



Area under ROC Curve : 0.644

We see the means of the independent variables overall and by value of the dependent variable. In this sample, there is a substantial difference between the mean *LWD* across birth weight groups but an apparently small *AGE* difference.

*AGE* is clearly not significant by conventional standards if we look at the coefficient/standard-error ratio. The confidence interval for the odds ratio (0.898, 1.019) includes 1.00, indicating no effect in relative risk, when adjusting for *LWD*. Before concluding that *AGE* does not belong in the model, H&L consider the interaction of *AGE* and *LWD*.

### Interpretation of the Fitted Model

Consider the *HOSLEM* data. Here we fit the model using *LWT* and *RACE* as independent variables.

The input is:

```
USE HOSLEM
LOGIT
CATEGORY RACE / DUMMY
MODEL LOW = CONSTANT + LWT + RACE
SAVE PREPROB
ESTIMATE
```

The output is:

#### Logistic Regression

Categorical values encountered during processing are

Variables	Levels
RACE (3 levels)	1.000 2.000 3.000
LOW (2 levels)	0.000 1.000

Categorical variables are dummy coded with the highest value as reference

#### Binary LOGIT Analysis

```
Dependent Variable : LOW
Input Records      : 189
Records for Analysis : 189
```

#### Sample Split

Category Choices	
0 (REFERENCE)	130
1 (RESPONSE)	59
Total	189

#### Log-Likelihood Iteration History

```
Log-Likelihood at Iteration1 : -131.005
Log-Likelihood at Iteration2 : -112.024
Log-Likelihood at Iteration3 : -111.632
Log-Likelihood at Iteration4 : -111.630
Log-Likelihood at Iteration5 : -111.630
Log-Likelihood                : -111.630
```

#### Information Criteria

```
AIC : 231.259
Schwarz's BIC : 244.226
```

## Logistic Regression

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	1.286	0.797	1.615	0.106	-0.275	2.848
2 LWT	-0.015	0.006	-2.364	0.018	-0.028	-0.003
3 RACE_1	-0.481	0.357	-1.347	0.178	-1.180	0.218
4 RACE_2	0.600	0.509	1.180	0.238	-0.397	1.598

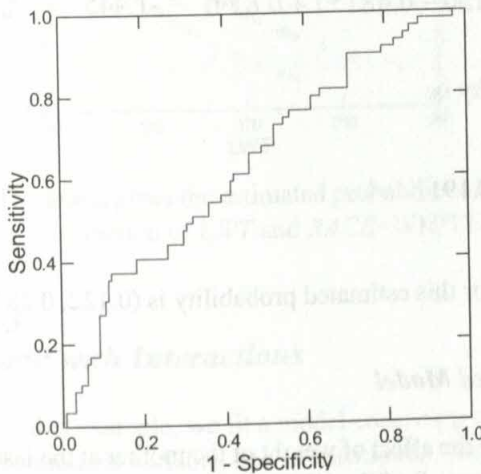
## Odds Ratio Estimates

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
2 LWT	0.985	0.006	0.973	0.997
3 RACE_1	0.618	0.221	0.307	1.244
4 RACE_2	1.823	0.928	0.672	4.943

Log-Likelihood of Constants only Model = LL(0) : -117.336  
 2\*[LL(N)-LL(0)] : 11.413  
 df : 3  
 p-value : 0.010

McFadden's Rho-squared : 0.049  
 Cox and Snell R-square : 0.059  
 Naglekerke's R-square : 0.082

## Receiver Operating Characteristic Curve



Area under ROC Curve : 0.648  
 SYSTAT save file created.  
 189 records written to SYSTAT save file.

From the *Parameter Estimates* table we get the estimated coefficients for the continuous variable *LWT* and the two dummy variables *RACE\_1* and *RACE\_2*. The estimates of the fitted values, logit and the standard error of the logit can be obtained in SYSTAT by giving the SAVE command prior to ESTIMATE command. The saved file *PREPROB* contains the estimated logits, standard error of the logits, predicted probabilities, upper and, lower bounds of the predicted probability.

The predicted probabilities are obtained from the following equation:

$$\hat{\pi}(x) = e^{g(x)} / (1 + e^{g(x)})$$

The estimated logit is obtained from the following equation:

$$\hat{g}(x) = 1.286 - 0.015 * LWT - 0.481 * RACE\_1 + 0.6 * Race\_2$$

Using the above equations we can obtain the estimated logit for a 150 pound white woman. The estimated logit is:

$$\hat{g}(x) = 1.286 - 0.015 * 150 - 0.481 * 1 + 0.6 * 0 = -1.445$$

And the estimated probability is:

$$\hat{\Pi}(x) = \frac{e^{-1.445}}{1 + e^{-1.445}} = 0.191$$

The 95% confidence interval for this estimated probability is (0.122, 0.285).

### ***Graphical presentation of the Fitted Model***

We can also present graphically the effect of weight of the mother at the last menstrual on birth weight taking into account *RACE* = WHITE as constant variable.



The input is:

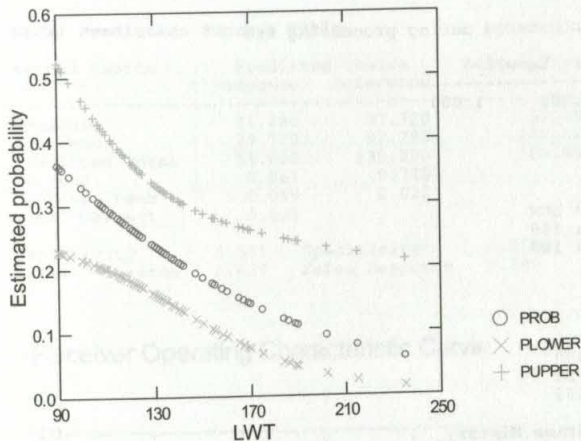
```

MERGE Hoslem preprob
SELECT (RACE =1)
PLOT PROB PLOWER PUPPER*LWT / OVERLAY,
YLABEL = 'Estimated probability',
xmin =90, xmax = 250

```

The output is:

Data for the following results were selected according to  
 SELECT (RACE =1)



The graph gives the estimated probability of low weight birth and the confidence band as the function of *LWT* and *RACE=WHITE*.

### Example 3

#### Binary Logit with Interactions

In this example, we fit a model consisting of a constant, a dummy variable, a continuous variable, and an interaction. Note that it is not necessary to create a new interaction variable; this is done for us automatically by writing the interaction on the MODEL statement. Let's also add a prediction table for this model.

The input is:

```
USE HOSLEM
LOGIT
MODEL LOW=CONSTANT+LWD+AGE+LWD*AGE
ESTIMATE / PREDICTION
SAVE SIM319/"SAVE ODDS RATIOS FOR H and L TABLE 3.19"
SIMULATE CONSTANT=0,AGE=0,LWD=1 / DO LWD*AGE =15,45,5
USE SIM319
LIST
```

The output is:

#### Logistic Regression

Categorical values encountered during processing are

Variables	Levels
LOW (2 levels)	0.000 1.000

Total : 12

Binary LOGIT Analysis

Dependent Variable : LOW  
Input Records : 189  
Records for Analysis : 189

#### Sample Split

Category Choices	
0 (REFERENCE)	130
1 (RESPONSE)	59
Total	189

#### Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-131.005
Log-Likelihood at Iteration2	-110.937
Log-Likelihood at Iteration3	-110.573
Log-Likelihood at Iteration4	-110.570
Log-Likelihood at Iteration5	-110.570
Log-Likelihood	-110.570

#### Information Criteria

AIC	229.140
Schwarz's BIC	242.107

#### Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	0.774	0.910	0.851	0.395	-1.009	2.558
2 LWD	-1.944	1.725	-1.127	0.260	-5.325	1.436
3 AGE	-0.080	0.040	-2.008	0.045	-0.157	-0.002
4 AGE*LWD	0.132	0.076	1.746	0.081	-0.016	0.281

**Odds Ratio Estimates**

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
2 LWD	0.143	0.247	0.005	4.206
3 AGE	0.924	0.037	0.854	0.998
4 AGE*LWD	1.141	0.086	0.984	1.324

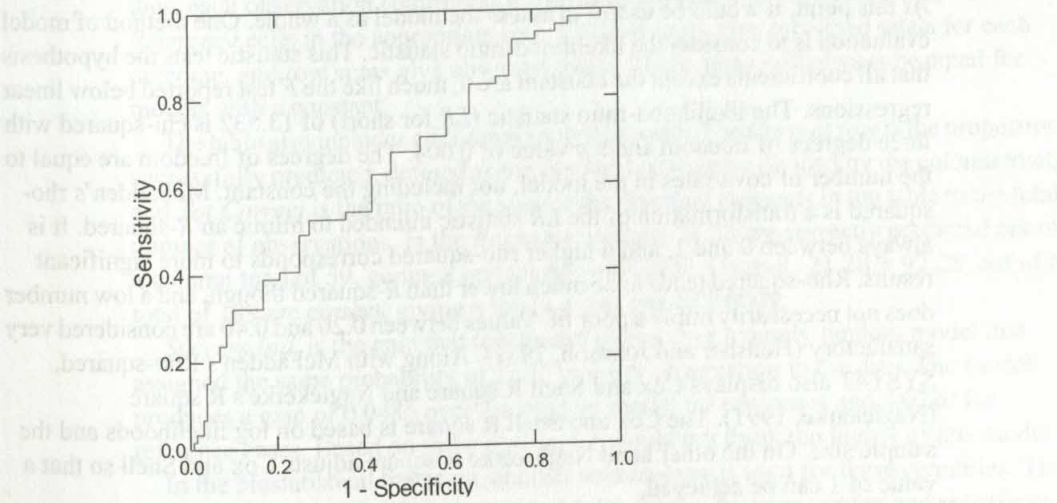
Log-Likelihood of Constants only Model = LL(0) : -117.336  
 2\*[LL(N)-LL(0)] : 13.532  
 df : 3  
 p-value : 0.004

McFadden's Rho-squared : 0.058  
 Cox and Snell R-square : 0.069  
 Naglekerke's R-square : 0.097

**Model Prediction Success Table**

Actual Choice	Predicted Choice		Actual Total
	Response	Reference	
Response	21.280	37.720	59.000
Reference	37.720	92.280	130.000
Predicted Total	59.000	130.000	189.000
Correct	0.361	0.710	
Success Index	0.049	0.022	
Total Correct	0.601		
Sensitivity	0.361	Specificity	0.710
False Reference	0.639	False Response	0.290

**Receiver Operating Characteristic Curve**



Area under ROC Curve : 0.659

**Logistic Regression: Simulation**  
**Simulation Vector**

Fixed Parameter	Value
1 CONSTANT	0.000
2 LWD	1.000
3 AGE	0.000

Loop Parameter	Maximum	Minimum	Increment
4 AGE*LWD	15.000	45.000	5.000

SYSTAT save file created.

7 records written to SYSTAT save file.

**List Cases**

Case	LOGIT ODDSL	SELOGIT ODDSU	PROB LOOP (1)	PLOWER	PUPPER	ODDS
1	0.039	0.660	0.510	0.222	0.791	1.040
	0.285	3.793	15.000			
2	0.700	0.404	0.668	0.477	0.816	2.013
	0.913	4.441	20.000			
3	1.361	0.420	0.796	0.631	0.899	3.899
	1.713	8.877	25.000			
4	2.022	0.690	0.883	0.661	0.967	7.552
	1.954	29.194	30.000			
5	2.683	1.031	0.936	0.660	0.991	14.626
	1.940	110.258	35.000			
6	3.344	1.391	0.966	0.650	0.998	28.326
	1.854	432.767	40.000			
7	4.005	1.759	0.982	0.636	0.999	54.859
	1.745	1724.151	45.000			

At this point, it would be useful to assess the model as a whole. One method of model evaluation is to consider the likelihood-ratio statistic. This statistic tests the hypothesis that all coefficients except the constant are 0, much like the  $F$  test reported below linear regressions. The likelihood-ratio statistic ( $LR$  for short) of 13.532 is chi-squared with three degrees of freedom and a  $p$  value of 0.004. The degrees of freedom are equal to the number of covariates in the model, not including the constant. McFadden's rho-squared is a transformation of the  $LR$  statistic intended to mimic an  $R$ -squared. It is always between 0 and 1, and a higher rho-squared corresponds to more significant results. Rho-squared tends to be much lower than  $R$ -squared though, and a low number does not necessarily imply a poor fit. Values between 0.20 and 0.40 are considered very satisfactory (Hensher and Johnson, 1981). Along with McFadden's Rho-squared, SYSTAT also displays Cox and Snell  $R$  square and Naglekerke's  $R$  square (Naglekerke, 1991). The Cox and Snell  $R$  square is based on log likelihoods and the sample size. On the other hand Naglekerke  $R$  square adjusts Cox and Snell so that a value of 1 can be achieved.



Models can also be assessed relative to one another. A likelihood-ratio test is formally conducted by computing twice the difference in log-likelihoods for any pair of nested models. Commonly called the  $G$  statistic, it has degrees of freedom equal to the difference in the number of parameters estimated in the two models. Comparing the current model with the model without the interaction, we have

$$G = 2 * (112.14338 - 110.56997) = 3.14684$$

with one degree of freedom, which has a  $p$  value of 0.076. This result corresponds to the bottom row of H&L's Table 3.17. The conclusion of the test is that the interaction approaches significance.

### **Prediction Success Table**

The output also includes a prediction success table, which summarizes the classificatory power of the model. The rows of the table show how observations from each level of the dependent variable are allocated to predicted outcomes. Reading across the first (*Response*) row we see that of the 59 cases of low birth weight, 21.28 are correctly predicted and 37.72 are incorrectly predicted. The second row shows that of the 130 not-*LOW* cases, 37.72 are incorrectly predicted and 92.28 are correctly predicted.

By default, the prediction success table sums predicted probabilities into each cell; thus, each observation contributes a fractional amount to both the *Response* and *Reference* cells in the appropriate row. Column sums give predicted totals for each outcome, and row sums give observed totals. These sums will always be equal for models with a constant.

The table also includes additional analytic results. The *Correct* row is the proportion successfully predicted, defined as the diagonal table entry divided by the column total, and *Tot.Correct* is the ratio of the sum of the diagonal elements in the table to the total number of observations. In the *Response* column, 21.28 are correctly predicted out of a column total of 59, giving a correct rate of 0.3607. Overall, 21.28 + 92.28 out of a total of 189 are correct, giving a total correct rate of 0.6009.

*Success Ind.* is the gain that this model shows over a purely random model that assigned the same probability of *LOW* to every observation in the data. The model produces a gain of 0.0485 over the random model for responses and 0.0220 for reference cases. Based on these results, we would not think too highly of this model.

In the biostatistical literature, another terminology is used for these quantities. The *Correct* quantity is also known as **sensitivity** for the *Response* group and **specificity**

for the *Reference* group. The *False Reference* rate is the fraction of those predicted to respond that actually did not respond, while the *False Response* rate is the fraction of those predicted to not respond that actually responded.

We prefer the prediction success terminology because it is applicable to the multinomial case as well.

### Simulation

To understand the implications of the interaction, we need to explore how the relative risk of low birth weight varies over the typical child-bearing years. This changing relative risk is evaluated by computing the logit difference for base and comparison groups. The logit for the base group, mothers with  $LWD = 0$ , is written as  $L(0)$ ; the logit for the comparison group, mothers with  $LWD = 1$ , is  $L(1)$ . Thus,

$$\begin{aligned} L(0) &= \text{CONSTANT} + B2*AGE \\ L(1) &= \text{CONSTANT} + B1*LWD + B2*AGE + B3*LWD*AGE \\ &= \text{CONSTANT} + B1 + B2*AGE + B3*AGE \end{aligned}$$

since, for  $L(1)$ ,  $LWD = 1$ . The logit difference is

$$L(1) - L(0) = B1 + B3*LWD*AGE$$

which is the coefficient on  $LWD$  plus the interaction multiplied by its coefficient. The difference  $L(1) - (0)$  evaluated for a mother of a given age is a measure of the log relative risk due to  $LWD$  being 1. This can be calculated simply for several ages, and converted to odds ratios with upper and lower confidence bounds, using the `SIMULATE` command.

`SIMULATE` calculates the predicted logit, predicted probability, odds ratio, upper and lower bounds, and the standard error of the logit for any specified values of the covariates. In the above command, the constant and age are set to 0, because these coefficients do not appear in the logit difference.  $LWD$  is set to 1, and the interaction is allowed to vary from 15 to 45 in increments of five years. The only printed output produced by this command is a summary report.

`SIMULATE` does not print results when a `DO LOOP` is specified because of the potentially large volume of output it can generate. To view the results, use the commands:

```
USE SIM319
LIST
```

The results give the effect of low maternal weight (*LWD*) on low birth weight as a function of age, where  $LOOP(1)$  is the value of  $AGE * LWD$  (which is just  $AGE$ ) and  $ODDSU$  and  $ODDSL$  are upper and lower bounds of the odds ratio. We see that the effect of *LWD* goes up dramatically with age, although the confidence interval becomes quite large beyond age 30. The results presented here are calculated internally within LOGIT and thus differ slightly from those reported in H&L, who use printed output with fewer decimal places of precision to obtain their results.

### Example 4

#### Deciles of Risk and Model Diagnostics

Before turning to more detailed model diagnostics, we fit H&L's final model. As a result of experimenting with more variables and a large number of interactions, H&L arrive at the model used here.

The input is:

```
USE HOSLEM
LOGIT
CATEGORY RACE / DUMMY
MODEL LOW=CONSTANT+AGE+RACE+SMOKE+HT+UI+LWD+PTD+,
        AGE*LWD+SMOKE*LWD
ESTIMATE
SAVE RESIDDC
DC / P=0.06850,0.09360,0.15320,0.20630,0.27810,0.33140,
      0.42300,0.49124,0.61146

USE RESIDDC
PPLOT PEARSON / SIZE=VARIANCE
PLOT DELPSTAT*PROB/SIZE=DELBETA(1)
```

The categorical variable *RACE* is specified to have three levels. By default LOGIT uses the highest category as the reference group, although this can be changed. The model includes all of the main variables except *FTV*, with *LWT* and *PTL* transformed into dummy variable variants *LWD* and *PTD*, and two interactions. To reproduce the results of Table 5.1 of H&L, we specify a particular set of cut points for the deciles of risk table.



The output is:

### Logistic Regression

Categorical values encountered during processing are

Variables	Levels		
RACE (3 levels)	1.000	2.000	3.000
LOW (2 levels)	0.000	1.000	

Categorical variables are dummy coded with the highest value as reference

### Binary LOGIT Analysis

Dependent Variable : LOW  
Input Records : 189  
Records for Analysis : 189

### Sample Split

Category Choices	
0 (REFERENCE)	130
1 (RESPONSE)	59
Total	189

### Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-131.005
Log-Likelihood at Iteration2	-98.066
Log-Likelihood at Iteration3	-96.096
Log-Likelihood at Iteration4	-96.006
Log-Likelihood at Iteration5	-96.006
Log-Likelihood at Iteration6	-96.006
Log-Likelihood	-96.006

### Information Criteria

AIC	214.012
Schwarz's BIC	249.672

### Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	0.248	1.068	0.232	0.816	-1.845	2.340
2 AGE	-0.084	0.046	-1.843	0.065	-0.173	0.005
3 RACE 1	-0.760	0.464	-1.637	0.102	-1.669	0.150
4 RACE 2	0.323	0.532	0.608	0.543	-0.719	1.366
5 SMOKE	1.153	0.458	2.515	0.012	0.255	2.052
6 HT	1.359	0.661	2.055	0.040	0.063	2.656
7 UI	0.728	0.479	1.519	0.129	-0.212	1.668
8 LWD	-1.730	1.868	-0.926	0.354	-5.392	1.932
9 PTD	1.232	0.471	2.613	0.009	0.308	2.155
10 AGE*LWD	0.147	0.083	1.779	0.075	-0.015	0.310
11 SMOKE*LWD	-1.407	0.819	-1.719	0.086	-3.012	0.197



## Odds Ratio Estimates

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
2 AGE	0.919	0.042	0.841	1.005
3 RACE 1	0.468	0.217	0.188	1.162
4 RACE 2	1.382	0.735	0.487	3.920
5 SMOKE	3.168	1.452	1.290	7.781
6 HT	3.893	2.575	1.065	14.235
7 UI	2.071	0.993	0.809	5.301
8 LWD	0.177	0.331	0.005	6.902
9 PTD	3.427	1.615	1.360	8.632
10 AGE*LWD	1.159	0.096	0.985	1.363
11 SMOKE*LWD	0.245	0.200	0.049	1.218

Log-Likelihood of Constants only Model = LL(0) : -117.336

2\*[LL(N)-LL(0)] : 42.660

df : 10

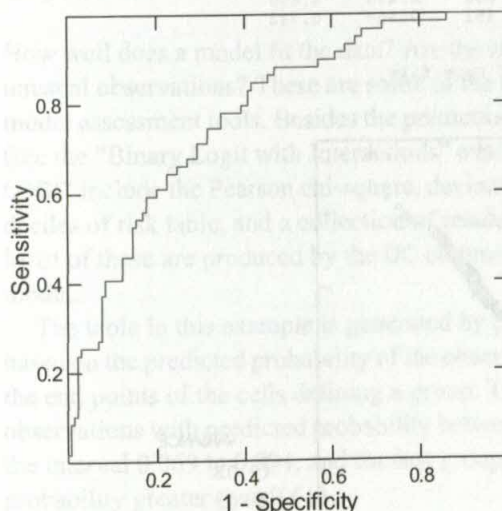
p-value : 0.000

McFadden's Rho-squared : 0.182

Cox and Snell R-square : 0.202

Naglerkerke's R-square : 0.284

## Receiver Operating Characteristic Curve



Area under ROC Curve : 0.785

### Logistic Regression: Deciles of Risk

Deciles of Risk

Records Processed : 189

Sum of weights : 189.000

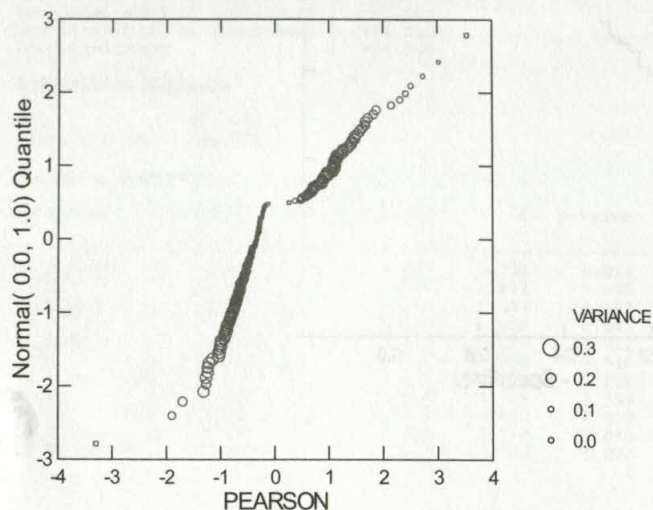
	Statistic	p-value	df
Hosmer-Lemeshow*	5.231	0.733	8.000
Pearson	183.443	0.374	178.000
Deviance	192.012	0.224	178.000

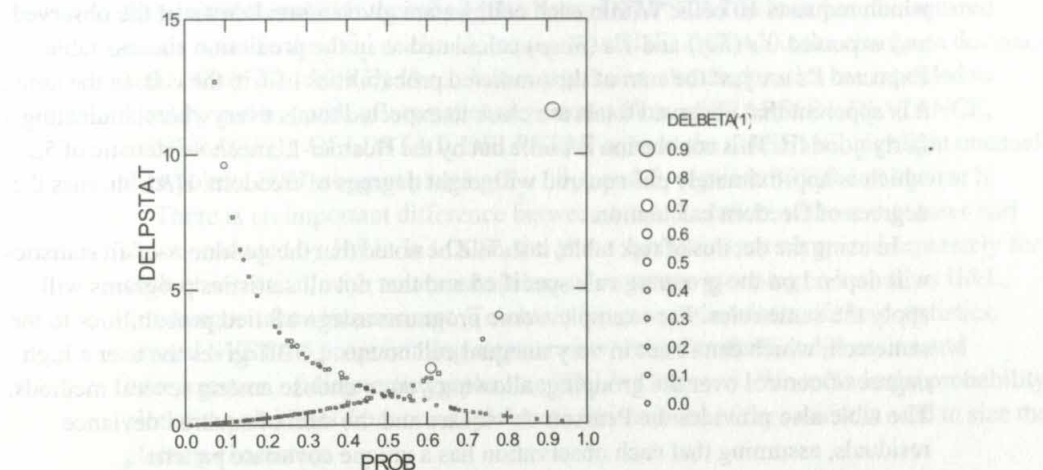
\* Large influence of one or more deciles may affect statistic.

Category	0.069	0.094	0.153	0.206	0.278	0.331	0.423
Response Observation	0.000	1.000	4.000	2.000	6.000	6.000	6.000
Expected Value	0.854	1.641	2.252	3.646	5.017	5.566	6.816
Reference Observation	18.000	19.000	14.000	18.000	14.000	12.000	12.000
Expected Value	17.146	18.359	15.748	16.354	14.983	12.434	11.184
Average Probability	0.047	0.082	0.125	0.182	0.251	0.309	0.379
Category	0.491	0.611	1.000				
Response Observation	10.000	9.000	15.000				
Expected Value	8.570	10.517	14.122				
Reference Observation	9.000	10.000	4.000				
Expected Value	10.430	8.483	4.878				
Average Probability	0.451	0.554	0.743				

SYSTAT save file created.

189 records written to SYSTAT save file.





### Deciles of Risk

How well does a model fit the data? Are the results unduly influenced by a handful of unusual observations? These are some of the questions we try to answer with our model assessment tools. Besides the prediction success table and likelihood-ratio tests (see the “Binary Logit with Interactions” example), the model assessment methods in LOGIT include the Pearson chi-square, deviance and Hosmer-Lemeshow statistics, the deciles of risk table, and a collection of residual, leverage, and influence quantities. Most of these are produced by the DC command, which is invoked after estimating a model.

The table in this example is generated by partitioning the sample into 10 groups based on the predicted probability of the observations. The row labeled *Category* gives the end points of the cells defining a group. Thus, the first group consists of all observations with predicted probability between 0 and 0.069, the second group covers the interval 0.069 to 0.094, and the last group contains observations with predicted probability greater than 0.611.

The cell end points can be specified explicitly as we did or generated automatically by LOGIT. Cells will be equally spaced if the DC command is given without any arguments, and LOGIT will allocate approximately equal numbers of observations to each cell when the BINS option is given, as:

```
DC / BINS = 10
```

which requests 10 cells. Within each cell, we are given a breakdown of the observed and expected 0's (*Ref*) and 1's (*Resp*) calculated as in the prediction success table. Expected 1's are just the sum of the predicted probabilities of 1 in the cell. In the table, it is apparent that observed totals are close to expected totals everywhere, indicating a fairly good fit. This conclusion is borne out by the Hosmer-Lemeshow statistic of 5.23, which is approximately chi-squared with eight degrees of freedom. H&L discuss the degrees of freedom calculation.

In using the deciles of risk table, it should be noted that the goodness-of-fit statistics will depend on the grouping rule specified and that not all statistics programs will apply the same rules. For example, some programs assign all tied probabilities to the same cell, which can result in very unequal cell counts. LOGIT gives the user a high degree of control over the grouping, allowing you to choose among several methods. The table also provides the Pearson chi-square and the sum of squared deviance residuals, assuming that each observation has a unique covariate pattern.

### Regression Diagnostics

If the DC command is preceded by a SAVE command, a SYSTAT data file containing regression diagnostics will be created (Pregibon, 1981; Cook and Weisberg, 1982). The SAVE file contains these variables:

ACTUAL	Value of Dependent Variable
PREDICT	Class Assignment (1 or 0)
PROB	Predicted probability
LEVERAGE(1)	Diagonal element of Pregibon "hat" matrix
LEVERAGE(2)	Component of <i>LEVERAGE(1)</i>
PEARSON	Pearson Residual for observation
VARIANCE	Variance of Pearson Residual
STANDARD	Standardized Pearson Residual
DEVIANCE	Deviance Residual
DELDSTART	Change in Deviance chi-square
DELPSTART	Change in Pearson chi-square
DELBETA(1)	Standardized Change in Beta
DELBETA(2)	Standardized Change in Beta
DELBETA(3)	Standardized Change in Beta

LEVERAGE(1) is a measure of the influence of an observation on the model fit and is H&L's *h*. DELBETA(1) is a measure of the change in the coefficient vector due to



the observation and is their  $\delta_\beta$  (delta beta), DELPSTAT is based on the squared residual and is their  $\delta_{\chi^2}$  (delta chi-square), and DELDSTAT is the change in deviance and is their  $\delta_D$  (delta D). As in linear regression, the diagnostics are intended to identify outliers and influential observations. Plots of PEARSON, DEVIANCE, LEVERAGE(I), DELDSTAT, DELPSTAT against the CASE will highlight unusual data points. H&L suggest plotting  $\delta_{\chi^2}$ ,  $\delta_D$ , and  $\delta_\beta$  against PROB and against h.

There is an important difference between our calculation of these measures and those produced by H&L. In LOGIT, the above quantities are computed separately for each observation, with no account taken of covariate grouping; whereas, in H&L, grouping is taken into account. To obtain the grouped variants of these statistics, several SYSTAT programming steps are involved. For further discussion and interpretation of diagnostic graphs, see H&L's Chapter 5. We include the probability plot of the residuals from our model, with the variance of the residuals used to size the plotting characters.

We also display an example of the graph on the cover of H&L. The original cover was plotted using SYSTAT Version 5 for the Macintosh. There are slight differences between the two plots because of the scales and number of iterations in the model fitting, but the examples are basically the same. H&L is an extremely valuable resource for learning about graphical aids to diagnosing logistic models.

### **Example 5** **Quantiles**

In bioassay, it is common to estimate the dosage required to kill 50% of a target population. For example, a toxicity experiment might establish the concentration of nicotine sulphate required to kill 50% of a group of common fruit flies (Hubert, 1984). More generally, the goal is to identify the level of a stimulus required to induce a 50% response rate, where the response is any binary outcome variable and the stimulus is a continuous covariate. In bioassay, stimuli include drugs, toxins, hormones, and insecticides; the responses include death, weight gain, bacterial growth, and color change, but the concepts are equally applicable to other sciences.

To obtain the LD50 in LOGIT, simply issue the QNTL command. However, don't make the mistake of spelling "quantile" as QU, which means QUIT in SYSTAT. QNTL will produce not only the LD50 but also a number of other quantiles as well, with upper

and lower bounds when they exist. Consider the following data *WILL* from Williams (1986):

	RESPONSE	LDOSE	COUNT
CASE 1	1	-2	1
CASE 2	0	-2	4
CASE 3	1	-1	3
CASE 4	0	-1	2
CASE 5	1	0	2
CASE 6	0	0	3
CASE 7	1	1	4
CASE 8	0	1	1
CASE 9	1	2	5

Here, *RESPONSE* is the dependent variable, *LDOSE* is the logarithm of the dose (stimulus), and *COUNT* is the number of subjects with that response.

The input is:

```
USE WILL
FREQ COUNT
LOGIT
MODEL RESPONSE=CONSTANT+LDOSE
ESTIMATE
QNTL
```

The output is:

#### Logistic Regression

Case frequencies determined by value of variable COUNT  
Categorical values encountered during processing are

Variables	Levels
RESPONSE (2 levels)	0.000 1.000

#### Binary LOGIT Analysis

```
Dependent Variable      : RESPONSE
Analysis is Weighted by : COUNT
Sum of Weights          : 25.000
Input Records           : 9
Records for Analysis    : 9
```

#### Sample Split

Category	Count	Weighted Count
0 (REFERENCE)	4	10.000
1 (RESPONSE)	5	15.000
Total	9	25.000

**Log-Likelihood Iteration History**

```

Log-Likelihood at Iteration1  : -17.329
Log-Likelihood at Iteration2  : -13.277
Log-Likelihood at Iteration3  : -13.114
Log-Likelihood at Iteration4  : -13.112
Log-Likelihood at Iteration5  : -13.112
Log-Likelihood                : -13.112

```

**Information Criteria**

```

AIC          : 30.224
Schwarz's BIC : 30.618

```

**Parameter Estimates**

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	0.564	0.496	1.138	0.255	-0.408	1.536
2 LDOSE	0.919	0.394	2.334	0.020	0.147	1.691

**Odds Ratio Estimates**

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
2 LDOSE	2.507	0.987	1.159	5.425

```

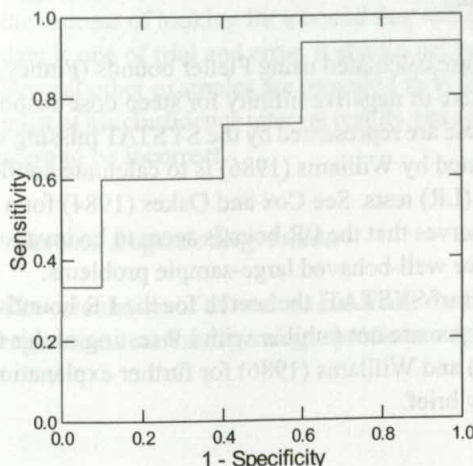
Log-Likelihood of Constants only Model = LL(0) : -16.825
2*[LL(N)-LL(0)] : 7.427
df : 1
p-value : 0.006

```

```

McFadden's Rho-squared : 0.221
Cox and Snell R-square : 0.562
Naglekerke's R-square : 0.576

```

**Receiver Operating Characteristic Curve**



Area under ROC Curve : 0.800

**Logistic Regression: Quantiles  
Evaluation Vector**

1 CONSTANT : 1.000  
2 LDOSE : VALUE

**Quantile Table**

Probability	LOGIT	LDOSE	Upper	Lower
0.999	6.907	6.900	44.788	3.518
0.995	5.293	5.145	33.873	2.536
0.990	4.595	4.385	29.157	2.105
0.975	3.664	3.372	22.875	1.519
0.950	2.944	2.590	18.042	1.050
0.900	2.197	1.777	13.053	0.530
0.750	1.099	0.582	5.928	-0.445
0.667	0.695	0.142	3.551	-1.047
0.500	0.000	-0.613	0.746	-3.364
0.333	-0.695	-1.369	-0.347	-7.392
0.250	-1.099	-1.809	-0.731	-9.987
0.100	-2.197	-3.004	-1.552	-17.266
0.050	-2.944	-3.817	-2.046	-22.281
0.025	-3.664	-4.599	-2.503	-27.126
0.010	-4.595	-5.612	-3.081	-33.416
0.005	-5.293	-6.372	-3.508	-38.136
0.001	-6.907	-8.127	-4.486	-49.055

This table includes LD (probability) values between 0.001 and 0.999. The median lethal *LDOSE* (log-dose) is -0.613 with upper and lower bounds of 0.746 and -3.364 for the default 95% confidence interval, corresponding to a dose of 0.542 with limits 2.11 and 0.0346.

### ***Indeterminate Confidence Intervals***

Quantile confidence intervals are calculated using Fieller bounds (Finney, 1978), which can easily include positive or negative infinity for steep dose-response relationships. In the output, these are represented by the SYSTAT missing value. If this happens, an alternative suggested by Williams (1986) is to calculate confidence bounds using likelihood-ratio (LR) tests. See Cox and Oakes (1984) for a likelihood profile example. Williams observes that the LR bounds seem to be invariably smaller than the Fieller bounds even for well-behaved large-sample problems.

With the BASIC commands of SYSTAT, the search for the LR bounds can be conducted easily. However, if you are not familiar with LR testing of this type, please refer to Cox and Oakes (1984) and Williams (1986) for further explanation, because our account here is necessarily brief.



We first estimate the model of *RESPONSE* on *LDOSE* reported above, which will be the unrestricted model in the series of tests. The key statistic is the final log-likelihood of  $-13.112$ . We then need to search for restricted models that force the LD50 to other values and that yield log-likelihoods no worse than  $-13.112 - 1.92 = -15.032$ . A difference in log-likelihoods of 1.92 marks a 95% confidence interval because  $2 * 1.92 = 3.84$  is the 0.95 cutoff of the chi-squared distribution with one degree of freedom.

A restricted model is estimated by using a new independent variable and fitting a model without a constant. The new independent variable is equal to the original minus the value of the hypothesized LD50 bound. Values of the bounds will be selected by trial and error.

Thus, to test an LD50 value of 0.4895, we could type:

```
LOGIT
LET LDOSEB=LDOSE-.4895
MODEL RESPONSE=LDOSEB
ESTIMATE
LET LDOSEB=LDOSE+2.634
MODEL RESPONSE=LDOSEB
ESTIMATE
```

The LET command is used to create the new variable *LDOSEB* “on the fly,” and the new model is then estimated without a constant. The only important part of the results from a restricted model is the final log-likelihood. It should be close to  $-15.032$  if we have found the boundary of the confidence interval. We won’t show the results of these estimations except to say that the lower bound was found to be  $-2.634$  and is tested using the second LET statement. Note that the value of the bound is subtracted from the original independent variable, resulting in the subtraction of a negative number. While the process of looking for a bound that will yield a log-likelihood of  $-15.032$  for these data is one of trial and error, it should not take long with the interactive program. Several other examples are provided in Williams (1986). We were able to reproduce most of his confidence interval results, but for several models his reported LD50 values seem to be incorrect.

### ***Quantiles and Logistic Regression***

The calculation of LD values has traditionally been conducted in the context of simple regressions containing a single predictor variable. LOGIT extends the notion to multiple

regression by allowing you to select one variable for LD calculations while holding the values of the other variables constant at prespecified values. Thus,

```
USE HOSLEM
CATEGORY RACE
MODEL LOW = CONSTANT + AGE + RACE + SMOKE + HT + ,
          UI + LWD + PTD
ESTIMATE
QNTL AGE / CONSTANT=1, RACE[1]=1, SMOKE=1, PTD=1,
          LWD=1, HT=1, UI=1
```

will produce the quantiles for *AGE* with the other variables set as specified. The Fieller bounds are calculated, adjusting for all other parameters estimated.

### Example 6

#### Multinomial Logit

We will illustrate multinomial modeling with an example, emphasizing what is new in this context. If you have not already read the example on binary logit, this is a good time to do so. The data used here have been extracted from the National Longitudinal Survey of Young Men, 1979. Information on 200 individuals is supplied on school enrollment status (*NOTENR* = 1 if not enrolled, 0 otherwise), log10 of wage (*LW*), age, highest completed grade (*EDUC*), mother's education (*MED*), father's education (*FED*), an index of reading material available in the home (*CULTURE* = 1 for least, 3 for most), mean income of persons in father's occupation in 1960 (*FOMY*), an IQ measure, a race dummy (*BLACK* = 0 for white), a region dummy (*SOUTH* = 0 for non-South), and the number of siblings (*NSIBS*).

We estimate a model to analyze the *CULTURE* variable, predicting its value with several demographic characteristics. In this example, we ignore the fact that the dependent variable is ordinal and treat it as a nominal variable. (See Agresti, 2002, for a discussion of the distinction.)

The input is:

```
USE NLS
FORMAT 4
PLENGTH LONG
LOGIT
MODEL CULTURE=CONSTANT+MED+FOMY
ESTIMATE / MEANS, PREDICT, CLASS, DERIVATIVE=INDIVIDUAL
PLENGTH
```

These commands look just like our binary logit analyses with the exception of the DERIVATIVE and CLASS options, which we will discuss below.

The output is:

### Logistic Regression

Categorical values encountered during processing are

Variables	Levels		
CULTURE (3 levels)	1.000	2.000	3.000

Total : 21

### Multinomial LOGIT Analysis

Dependent Variable : CULTURE  
Input Records : 200  
Records for Analysis : 200

### Sample Split

Category Choices	
1	12
2	49
3 (REFERENCE)	139
Total	200

### Independent Variable Means

PARAMETER	1	2	3	OVERALL
1 CONSTANT	1.0000	1.0000	1.0000	1.0000
2 MED	8.7500	10.1837	11.4460	10.9750
3 FOMY	4551.5000	5368.8571	6116.1367	5839.1750

### Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-219.7225
Log-Likelihood at Iteration2	-145.2936
Log-Likelihood at Iteration3	-138.9952
Log-Likelihood at Iteration4	-137.8612
Log-Likelihood at Iteration5	-137.7851
Log-Likelihood at Iteration6	-137.7846
Log-Likelihood at Iteration7	-137.7846
Log-Likelihood	-137.7846

### Information Criteria

AIC : 287.5692  
Schwarz's BIC : 307.3591

### Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
Choice Group: 1						
1 CONSTANT	5.0638	1.6964	2.9850	0.0028	1.7389	8.3886
2 MED	-0.4228	0.1423	-2.9711	0.0030	-0.7017	-0.1439
3 FOMY	-0.0006	0.0002	-2.6034	0.0092	-0.0011	-0.0002
Choice Group: 2						
1 CONSTANT	2.5435	0.9834	2.5864	0.0097	0.6161	4.4709
2 MED	-0.1917	0.0768	-2.4956	0.0126	-0.3423	-0.0411
3 FOMY	-0.0003	0.0001	-2.1884	0.0286	-0.0005	0.0000



## Odds Ratio Estimates

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
Choice Group: 1				
2 MED	0.6552	0.0932	0.4958	0.8660
3 FOMY	0.9994	0.0002	0.9989	0.9998
Choice Group: 2				
2 MED	0.8255	0.0634	0.7101	0.9597
3 FOMY	0.9997	0.0001	0.9995	1.0000

Log-Likelihood of Constants only Model = LL(0) : -153.2535

2\*[LL(N)-LL(0)] : 30.9379

df : 4

p-value : 0.0000

McFadden's Rho-squared : 0.1009

Cox and Snell R-square : 0.1433

Naglekerke's R-square : 0.1828

## Wald Tests on Effects Across all Choices

Effect	Wald Statistic	Chi-square Significance	df
1 CONSTANT	12.0028	0.0025	2.0000
2 MED	12.1407	0.0023	2.0000
3 FOMY	9.4575	0.0088	2.0000

## Covariance Matrix

	1	2	3	4	5	6
1	2.8777					
2	-0.1746	0.0202				
3	-0.0002	0.0000	0.0000			
4	0.5097	-0.0282	0.0000	0.9670		
5	-0.0274	0.0027	0.0000	-0.0541	0.0059	
6	0.0000	0.0000	0.0000	-0.0001	0.0000	0.0000

## Correlation Matrix

	1	2	3	4	5	6
1	1.0000	-0.7234	-0.6151	0.3055	-0.2100	-0.1659
2	-0.7234	1.0000	-0.0633	-0.2017	0.2462	-0.0149
3	-0.6151	-0.0633	1.0000	-0.1515	-0.0148	0.2284
4	0.3055	-0.2017	-0.1515	1.0000	-0.7164	-0.5544
5	-0.2100	0.2462	-0.0148	-0.7164	1.0000	-0.1570
6	-0.1659	-0.0149	0.2284	-0.5544	-0.1570	1.0000

## Individual variable derivatives averaged over all observations.

PARAMETER	1	2	3
1 CONSTANT	0.2033	0.3441	-0.5474
2 MED	-0.0174	-0.0251	0.0425
3 FOMY	0.0000	0.0000	0.0001



Model Prediction Success Table

Actual Choice	Predicted Choice			Actual Total
	1	2	3	
1	1.8761	4.0901	6.0338	12.0000
2	3.6373	13.8826	31.4801	49.0000
3	6.4865	31.0273	101.4862	139.0000
Predicted Total	12.0000	49.0000	139.0000	200.0000
Correct	0.1563	0.2833	0.7301	
Success Index	0.0963	0.0383	0.0351	
Total Correct	0.5862			

Model Classification Table

Actual Choice	Predicted Choice			Actual Total
	1	2	3	
1	1.0000	3.0000	8.0000	12.0000
2	0.0000	4.0000	45.0000	49.0000
3	1.0000	5.0000	133.0000	139.0000
Predicted Total	2.0000	12.0000	186.0000	200.0000
Correct	0.0833	0.0816	0.9568	
Success Index	0.0233	-0.1634	0.2618	
Total Correct	0.6900			

The output begins with a report on the number of records read and retained for analysis. This is followed by a frequency table of the dependent variable; both weighted and unweighted counts would be provided if the FREQ option had been used. The means table provides means of the independent variables by value of the dependent variable. We observe that the highest educational and income values are associated with the most reading material in the home. Next, an abbreviated history of the optimization process lists the log-likelihood at each iteration, and finally, the estimation results are printed.

Note that the regression results consist of two sets of estimates, labeled *Choice Group 1* and *Choice Group 2*. It is this multiplicity of parameter estimates that differentiates multinomial from binary logit. If there had been five categories in the dependent variable, there would have been four sets of estimates, and so on. This volume of output provides the challenge to understanding the results.

The results are a little more intelligible when you realize that we have really estimated a series of binary logits simultaneously. The first submodel consists of the two dependent variable categories 1 and 3, and the second consists of categories 2 and 3. These submodels always include the highest level of the dependent variable as the reference class and one other level as the response class. If NCAT had been set to 25, the 24 submodels would be categories 1 and 25, categories 2 and 25, through categories 24 and 25. We then obtain the odds ratios for the two submodels separately, comparing dependent variable levels 1 against 3 and 2 against 3. This table shows that levels 1 and 2 are less likely as *MED* and *FOMY* increase, as the odds ratio is less than 1.

### **Wald Test Table**

The coefficient/standard-error ratios ( $z$  ratios) reported next to each coefficient are a guide to the significance of an individual parameter. But when the number of categories is greater than two, each variable corresponds to more than one parameter. The Wald test table automatically conducts the hypothesis test of dropping all parameters associated with a variable, and the degrees of freedom indicates how many parameters were involved. Because each variable in this example generates two coefficients, the Wald tests have two degrees of freedom each. Given the high individual  $z$  ratios, it is not surprising that every variable is also significant overall. The `PLENGTH LONG` option also produces the parameter covariance and correlation matrices.

### **Derivative Tables**

In a multinomial context, we will want to know how the probabilities of each of the outcomes will change in response to a change in the covariate values. This information is provided in the derivative table, which tells us, for example, that when *MED* increases by one unit, the probability of category 3 goes up by 0.042, and categories 1 and 2 go down by 0.017 and 0.025, respectively. To assess properly the effect of father's income, the variable should be rescaled to hundreds or thousands of dollars (or the *FORMAT* increased) because the effect of an increase of one dollar is very small. The sum of the entries in each row is always 0 because an increase in probability in one category must come about by a compensating decrease in other categories. There is no useful interpretation of the *CONSTANT* row.

In general, the table shows how probability is reallocated across the possible values of the dependent variable as the independent variable changes. It thus provides a global view of covariate effects that is not easily seen when considering each binary submodel separately. In fact, the overall effect of a covariate on the probability of an outcome can be of the opposite sign of its coefficient estimate in the corresponding submodel. This is because the submodel concerns only two of the outcomes, whereas the derivative table considers all outcomes at once.

This table was generated by evaluating the derivatives separately for each individual observation in the data set and then computing the mean; this is the theoretically correct way to obtain the results. A quick alternative is to evaluate the derivatives once at the sample average of the covariates. This method saves time (but at the possible cost of accuracy) and is requested with the option `DERIVATIVE=AVERAGE`.



### Prediction Success

The PREDICT option instructs LOGIT to produce the prediction success table, which we have already seen in the binary logit. (See Hensher and Johnson, 1981; McFadden, 1979.) The table will break down the distribution of predicted outcomes by actual choice, with diagonals representing correct predictions and off-diagonals representing incorrect predictions. For the multinomial model, the table will have dimensions NCAT by NCAT with additional marginal results. For our example model, the core table is 3 by 3.

Each row of the table takes all cases having a specific value of the dependent variable and shows how the model allocates those cases across the possible outcomes. Thus in row 1, the 12 cases that actually had *CULTURE* = 1 were distributed by the predictive model as 1.88 to *CULTURE* = 1, 4.09 to *CULTURE* = 2, and 6.03 to *CULTURE* = 3. These numbers are obtained by summing the predicted probability of being in each category across all of the cases with *CULTURE* actually equal to 1. A similar allocation is provided for every value of the dependent variable.

The prediction success table is also bordered by additional information—row totals are observed sums, and column totals are predicted sums and will be equal for any model containing a constant. The *Correct* row gives the ratio of the number correctly predicted in a column to the column total. Thus, among cases for which *CULTURE* = 1, the fraction correct is  $1.8761/12 = 0.1563$ ; for *CULTURE* = 3, the ratio is  $101.4862/139 = 0.7301$ . The total correct gives the fraction correctly predicted overall and is computed as the sum *Correct* in each column divided by the table total. This is  $(1.8761 + 13.8826 + 101.4862)/200 = 0.5862$ .

The success index measures the gain that the model exhibits in number correctly predicted in each column over a purely random model (a model with just a constant). A purely random model would assign the same probabilities of the three outcomes to each case, as illustrated below:

#### Random Probability Model Predicted Sample Fraction

PROB (*CULTURE*=1)=  $12/200 = 0.0600$   
 PROB (*CULTURE*=2)=  $49/200 = 0.2450$   
 PROB (*CULTURE*=3)=  $139/200 = 0.6950$

#### Success Index = CORRECT - Random Predicted

$0.1563 - 0.0600 = 0.0963$   
 $0.2833 - 0.2450 = 0.0383$   
 $0.7301 - 0.6950 = 0.0351$

Thus, the smaller the success index in each column, the poorer the performance of the model; in fact, the index can even be negative.

Normally, one prediction success table is produced for each model estimated. However, if the data have been separated into learning and test subsamples with BY, a

separate prediction success table will be produced for each portion of the data. This can provide a clear picture of the strengths and weaknesses of the model when applied to fresh data.

### ***Classification Tables***

Classification tables are similar to prediction success tables except that predicted choices instead of predicted probabilities are added into the table. Predicted choice is the choice with the highest probability. Mathematically, the classification table is a prediction success table with the predicted probabilities changed, setting the highest probability of each case to 1 and the other probabilities to 0.

In the absence of fractional case weighting, each cell of the main table will contain an integer instead of a real number. All other quantities are computed as they would be for the prediction success table. In our judgment, the classification table is not as good a diagnostic tool as the prediction success table. The option is included primarily for the binary logit to provide comparability with results reported in the literature.

### ***Example 7*** ***Conditional Logistic Regression***

Data must be organized in a specific way for the conditional logistic model; fortunately, this organization is natural for matched sample case-control studies. First, matched samples must be grouped together; all subjects from a given stratum must be contiguous. It is thus advisable to provide each set with a unique stratum number to facilitate the sorting and tracking of records. Second, the dependent variable gives the relative position of the case within a matched set. Thus, the dependent variable will be an integer between 1 and NCAT, and if the case is first in each stratum, then the dependent variable will be equal to 1 for every record in the data set.

To illustrate how to set up conditional logit models, we use data discussed at length by Breslow and Day (1980) on cases of endometrial cancer in a retirement community near Los Angeles. The data are reproduced in their Appendix III and are identified in SYSTAT as *MACK*.

The data set includes the dependent variable *CANCER*, the exposure variables *AGE*, *GALL* (gall bladder disease), *HYP* (hypertension), *OBESE*, *ESTROGEN*, *DOSE*, *DUR* (duration of conjugated estrogen exposure), *NON* (other drugs), some transformations of these variables, and a set identification number. The data are organized by sets, with



the case coming first, followed by four controls, and so on, for a total of 315 observations ( $63 * (4 + 1)$ ).

To estimate a model of the relative risks of gall bladder disease, estrogen use, and their interaction, you may proceed as follows:

```
USE MACK
PLENGTH LONG
LOGIT
MODEL DEPVAR=GALL+EST+GALL*EST ;
ALT SETSIZE
NCAT 5
ESTIMATE
```

There are three key points to notice about this sequence of commands. First, the NCAT command is required to let LOGIT know how many subjects there are in a matched set. Unlike the unconditional binary LOGIT, a unit of information in matched samples will typically span more than one line of data, and NCAT will establish the minimum size of each matched set. If each set contains the same number of subjects, the NCAT command completely describes the data organization. If there were a varying number of controls per set, the size of each set would be signaled with the ALT command together with the NCAT command specifying the maximum size of each match set, as in

```
NCAT 5
ALT SETSIZE
```

Here, *SETSIZE* is a variable containing the total number of subjects (number of controls plus 1) per set. Each set could have its own value.

The second point is that the matched set conditional logit never contains a constant; the constant is eliminated along with all other variables that do not vary among members of a matched set. The third point is the appearance of the semicolon at the end of the model. This is required to distinguish the conditional from the unconditional model.

After you specify the commands, the output produced includes:

The output is:

#### Logistic Regression

Conditional LOGIT, data organized by matched set.  
Categorical values encountered during processing are

Variables	Levels
DEPVAR (1 levels)	1.0000

## Conditional LOGIT Analysis

Dependent Variable : DEPVAR  
 Number of Alternatives : SETSIZE  
 Input Records : 315  
 Matched Sets for Analysis : 63

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1 : -101.395  
 Log-Likelihood at Iteration2 : -79.055  
 Log-Likelihood at Iteration3 : -76.887  
 Log-Likelihood at Iteration4 : -76.733  
 Log-Likelihood at Iteration5 : -76.731  
 Log-Likelihood at Iteration6 : -76.731  
 Log-Likelihood : -76.731

## Information Criteria

AIC : 159.461  
 Schwarz's BIC : 165.891

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval Lower
1 GALL	2.894	0.883	3.278	0.001	1.164
2 EST	2.700	0.612	4.414	0.000	1.501
3 GALL*EST	-2.053	0.995	-2.063	0.039	-4.003

## Parameter Estimates (contd...)

Parameter	Upper
1 GALL	4.625
2 EST	3.899
3 GALL*EST	-0.103

## Odds Ratio Estimates

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval Lower	Upper
1 GALL	18.072	15.958	3.201	102.013
2 EST	14.882	9.104	4.487	49.362
3 GALL*EST	0.128	0.128	0.018	0.902

## Covariance Matrix

	1	2	3
1	0.780		
2	0.340	0.374	
3	-0.784	-0.367	0.990

## Correlation Matrix

	1	2	3
1	1.000	0.629	-0.892
2	0.629	1.000	-0.602
3	-0.892	-0.602	1.000

The output begins with a report on the number of SYSTAT records read and the number of matched sets kept for analysis. The remaining output parallels the results produced by the unconditional logit model. The parameters estimated are coefficients of a linear

logit, the relative risks are derived by exponentiation, and the interpretation of the model is unchanged. Model selection will proceed as it would in linear regression; you might experiment with logarithmic transformations of the data, explore quadratic and higher-order polynomials in the risk factors, and look for interactions. Examples of such explorations appear in Breslow and Day (1980).

### *Varying Controls per set*

The following is an example of the conditional logistic regression for varying controls per set. The data used is a subset of SYSTAT data *HOSLEM*. For making this data suitable for the desired analysis we have omitted some cases and created four new variables *SETSIZE*, *GROUP*, *REC* and *DEPVAR* along the lines of the previous analysis. The mother's age (*AGE*) is used as the matching variable and low infant birth weight (*LOW*) is used for deciding case and controls.

The input is:

```
USE HOSLEMM
LOGIT
NCAT 14
ALT SETSIZE
MODEL DEPVAR = LWT + SMOKE + HT + UI ;
ESTIMATE
```

The output is:

#### **Logistic Regression**

Conditional LOGIT, data organized by matched set.  
Categorical values encountered during processing are

Variables	Levels
DEPVAR (1 levels)	1.00

#### **Conditional LOGIT Analysis**

```
Dependent Variable      : DEPVAR
Number of Alternatives  : SETSIZE
Input Records           : 137
Matched Sets for Analysis : 17
```

#### **Log-Likelihood Iteration History**

```
Log-Likelihood at Iteration1 : -34.196
Log-Likelihood at Iteration2 : -30.170
Log-Likelihood at Iteration3 : -30.130
Log-Likelihood at Iteration4 : -30.130
Log-Likelihood at Iteration5 : -30.130
Log-Likelihood               : -30.130
```

**Information Criteria**

AIC | 68.259  
 Schwarz's BIC | 71.592

**Parameter Estimates**

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval Lower
1 LWT	-0.001	0.009	-0.069	0.945	-0.018
2 SMOKE	1.076	0.558	1.928	0.054	-0.018
3 HT	1.394	1.284	1.086	0.278	-1.122
4 UI	1.585	0.736	2.155	0.031	0.144

**Parameter Estimates (contd...)**

Parameter	Upper
1 LWT	0.017
2 SMOKE	2.170
3 HT	3.910
4 UI	3.027

**Odds Ratio Estimates**

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval Lower	Upper
1 LWT	0.999	0.009	0.982	1.017
2 SMOKE	2.933	1.637	0.982	8.757
3 HT	4.030	5.173	0.326	49.880
4 UI	4.881	3.590	1.154	20.633

**Covariance Matrix**

	1	2	3	4
1	0.000			
2	0.000	0.311		
3	-0.003	0.101	1.648	
4	0.001	0.104	0.180	0.541

**Correlation Matrix**

	1	2	3	4
1	1.000	0.098	-0.275	0.187
2	0.098	1.000	0.141	0.252
3	-0.275	0.141	1.000	0.190
4	0.187	0.252	0.190	1.000

**Example 8****Discrete Choice Models**

The *CHOICE* data set contains hypothetical data motivated by McFadden (1979). The *CHOICE* variable represents which of the three transportation alternatives (*AUTO*, *POOL*, *TRAIN*) each subject prefers. The first subscripted variable in each choice category represents *TIME* and the second, *COST*. Finally, *SEX* represents the gender of the chooser, and *AGE*, the age.



A basic discrete choice model is estimated with:

```
USE CHOICE
LOGIT
SET TIME = AUTO(1), POOL(1), TRAIN(1)
SET COST = AUTO(2), POOL(2), TRAIN(2)
MODEL CHOICE=TIME+COST
ESTIMATE
```

There are two new features of this program. First, the word *TIME* is not a SYSTAT variable name; rather, it is a label we chose to remind us of time spent commuting. The group of names in the SET statement are valid SYSTAT variables corresponding, in order, to the three modes of transportation. Although there are three variable names in the SET variable, only one attribute is being measured.

The output is:

**Logistic Regression**  
Linear Restriction System  
Discrete Choice Models

Categorical values encountered during processing are

Variables	Levels
CHOICE (3 levels)	1.000 2.000 3.000

Discrete Choice Analysis

Dependent Variable : CHOICE  
Input Records : 29  
Records for Analysis : 29

**Sample Split**

Category	Choices
1	15
2	6
3 (REFERENCE)	8
Total	29

**Log-Likelihood Iteration History**

Log-Likelihood at Iteration1	-31.860
Log-Likelihood at Iteration2	-31.142
Log-Likelihood at Iteration3	-31.141
Log-Likelihood at Iteration4	-31.141
Log-Likelihood	-31.141

**Information Criteria**

AIC	66.282
Schwarz's BIC	69.017

**Parameter Estimates**

Parameter	Estimate	Standard Error	Z	p-value
1 TIME	-0.020	0.017	-1.169	0.243
2 COST	-0.088	0.145	-0.611	0.541

**Odds Ratio Estimates**

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
1 TIME	0.980	0.017	0.947	1.014
2 COST	0.915	0.133	0.689	1.216

**Covariance Matrix**

	1	2
1	0.000	
2	0.001	0.021

**Correlation Matrix**

	1	2
1	1.000	0.384
2	0.384	1.000

The output begins with a frequency distribution of the dependent variable and a brief iteration history and prints standard regression results for the parameters estimated.

A key difference between a conditional variable clause and a standard SYSTAT polytomous variable is that each clause corresponds to only one estimated parameter regardless of the value of NCAT, while each free-standing polytomous variable generates NCAT - 1 parameters. The difference is best seen in a model that mixes both types of variables (see Hoffman and Duncan, 1988, or Steinberg, 1987) for further discussion).

**Mixed Parameters**

The following is an example of mixing polytomous and conditional variables:

```
USE CHOICE
LOGIT
CATEGORY SEX$
SET TIME = AUTO(1), POOL(1), TRAIN(1)
SET COST = AUTO(2), POOL(2), TRAIN(2)
MODEL CHOICE=TIME+COST+SEX$+AGE
ESTIMATE
```

The hybrid model generates a single coefficient each for *TIME* and *COST* and two sets of parameters for the polytomous variables.

**The output is:****Logistic Regression**

Linear Restriction System

Discrete Choice Models

Categorical values encountered during processing are

Variables	Levels		
SEX\$ (2 levels)	Female	Male	
CHOICE (3 levels)	1.000	2.000	3.000

Categorical variables are effects coded with the highest value as reference. Effects coding is in force for the categorical independent variables in your model. Parameters and odds ratios are easier to interpret for dummy coded categoricals. Unless you have specific reasons for requesting effects coding, we suggest that you re-issue the category statement with the /dummy option and re-fit your model. See Hosmer & Lemeshow, for more information.

**Discrete Choice Analysis**

Dependent Variable : CHOICE  
 Input Records : 29  
 Records for Analysis : 29

**Sample Split**

Category Choices	
1	15
2	6
3 (REFERENCE)	8
Total	29

**Log-Likelihood Iteration History**

Log-Likelihood at Iteration1 : -31.860  
 Log-Likelihood at Iteration2 : -28.495  
 Log-Likelihood at Iteration3 : -28.477  
 Log-Likelihood at Iteration4 : -28.477  
 Log-Likelihood at Iteration5 : -28.477  
 Log-Likelihood : -28.477

**Information Criteria**

AIC : 68.955  
 Schwarz's BIC : 77.159

**Parameter Estimates**

Parameter	Estimate	Standard Error	Z	p-value
1 TIME	-0.018	0.020	-0.887	0.375
2 COST	-0.351	0.217	-1.615	0.106
Choice Group: 1				
3 SEX\$ _Female	0.328	0.509	0.645	0.519
4 AGE	0.026	0.014	1.850	0.064
Choice Group: 2				
3 SEX\$ _Female	0.024	0.598	0.040	0.968
4 AGE	-0.008	0.016	-0.500	0.617

**Odds Ratio Estimates**

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
1 TIME	0.982	0.020	0.945	1.022
2 COST	0.704	0.153	0.460	1.078

Choice Group: 1				
3 SEX\$ _Female	1.388	0.707	0.512	3.764
4 AGE	1.026	0.014	0.998	1.054
Choice Group: 2				
3 SEX\$ _Female	1.024	0.613	0.317	3.308
4 AGE	0.992	0.016	0.961	1.024

**Wald Tests on Effects Across all Choices**

Effect	Wald Statistic	Chi-square Significance	df
3 SEX\$ _Female	0.551	0.759	2.000
4 AGE	4.475	0.107	2.000

**Covariance Matrix**

	1	2	3	4	5	6
1	0.000					
2	0.001	0.047				
3	0.002	0.009	0.259			
4	0.000	-0.001	0.002	0.000		
5	0.002	-0.018	0.165	0.002	0.358	
6	0.000	0.001	0.002	0.000	0.003	0.000

**Correlation Matrix**

	1	2	3	4	5	6
1	1.000	0.180	0.150	-0.076	0.146	-0.266
2	0.180	1.000	0.084	-0.499	-0.140	0.310
3	0.150	0.084	1.000	0.230	0.543	0.193
4	-0.076	-0.499	0.230	1.000	0.281	0.265
5	0.146	-0.140	0.543	0.281	1.000	0.323
6	-0.266	0.310	0.193	0.265	0.323	1.000

**Varying Alternatives**

For some discrete choice problems, the number of alternatives available varies across choosers. For example, health researchers studying hospital choice pooled data from several cities in which each city had a different number of hospitals in the choice set (Luft et al., 1988). Transportation research may pool data from locations having train service with locations without trains. Carson, Hanemann, and Steinberg (1990) pool responses from two contingent valuation survey questions having differing numbers of alternatives. To let LOGIT know about this, there are two ways of proceeding. The most flexible is to organize the data by choice. With the standard data layout, use the ALT command, as in

```
ALT NCHOICES
```

where *NCHOICES* is a SYSTAT variable containing the number of alternatives available to the chooser. If the value of the ALT variable is less than NCAT for an observation, LOGIT will use only the first *NCHOICES* variables in each conditional variable clause in the analysis.



With the standard data layout, the ALT command is useful only if the choices not available to some cases all appear at the end of the choice list. Organizing data by choice is much more manageable. One final note on varying numbers of alternatives: if the ALT command is used in the standard data layout, the model may not contain a constant or any polytomous variables; the model must be composed only of conditional variable clauses. We will not show an example here because by now you must have figured that we believe the by-choice layout is more suitable if you have data with varying choice alternatives.

### Interactions

A common practice in discrete choice models is to enter characteristics of choosers as interactions with attributes of the alternatives in conditional variable clauses. When dealing with large sets of alternatives, such as automobile purchase choices or hospital choices, where the model may contain up to 60 different alternatives, adding polytomous variables can quickly produce unmanageable estimation problems, even for mainframes. In the transportation literature, it has become commonplace to introduce demographic variables as interactions with, or other functions, of the discrete choice variables. Thus, instead of, or in addition to, the *COST* group of variables, *AUTO(2)*, *POOL(2)*, *TRAIN(2)*, you might see the ratio of cost to income. These ratios would be created with LET transformations and then added in another SET list for use as a conditional variable in the MODEL statement. Interactions can also be introduced this way. By confining demographic variables to appear only as interactions with choice variables, the number of parameters estimated can be kept quite small.

Thus, an investigator might prefer

```
USE CHOICE
LOGIT
SET TIME = AUTO(1), POOL(1), TRAIN(1)
SET TIMEAGE=AUTO(1)*AGE, POOL(1)*AGE, TRAIN(1)*AGE
SET COST = AUTO(2), POOL(2), TRAIN(2)
MODEL CHOICE=TIME+TIMEAGE+COST
ESTIMATE
```

as a way of entering demographics. The advantage to using only conditional clauses is clear when dealing with a large value of NCAT as the number of additional parameters estimated is minimized. The model above yields:

## Logistic Regression

## Linear Restriction System

## Discrete Choice Models

Categorical values encountered during processing are

Variables	Levels		
CHOICE (3 levels)	1.000	2.000	3.000

## Discrete Choice Analysis

Dependent Variable : CHOICE  
 Input Records : 29  
 Records for Analysis : 29

## Sample Split

Category Choices	
1	15
2	6
3 (REFERENCE)	8
Total	29

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1 : -31.860  
 Log-Likelihood at Iteration2 : -28.021  
 Log-Likelihood at Iteration3 : -27.866  
 Log-Likelihood at Iteration4 : -27.864  
 Log-Likelihood at Iteration5 : -27.864  
 Log-Likelihood : -27.864

## Information Criteria

AIC : 61.728  
 Schwarz's BIC : 65.830

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value
1 TIME	-0.148	0.062	-2.382	0.017
2 TIMEAGE	0.003	0.001	2.193	0.028
3 COST	0.007	0.155	0.043	0.966

## Odds Ratio Estimates

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
1 TIME	0.863	0.054	0.764	0.974
2 TIMEAGE	1.003	0.001	1.000	1.006
3 COST	1.007	0.156	0.742	1.365

## Covariance Matrix

	1	2	3
1	0.004		
2	0.000	0.000	
3	-0.001	0.000	0.024

## Correlation Matrix

	1	2	3
1	1.000	-0.936	-0.110
2	-0.936	1.000	0.273
3	-0.110	0.273	1.000

**Constants**

The models estimated here deliberately did not include a constant because the constant is treated as a polytomous variable in LOGIT. To obtain an alternative specific constant, enter the following model statement:

```
USE CHOICE
LOGIT
SET TIME = AUTO(1), POOL(1), TRAIN(1)
SET COST = AUTO(2), POOL(2), TRAIN(2)
MODEL CHOICE=CONSTANT+TIME+COST
ESTIMATE
```

Two CONSTANT parameters would be estimated. For the discrete choice model with the type of data layout of this example, there is no need to specify the NCAT value because LOGIT determines this automatically by the number of variables between the brackets. If the model statement is inconsistent in the number of variables within brackets across conditional variable clauses, an error message will be generated.

The output is:

Logistic Regression

Linear Restriction System  
Discrete Choice Models

Categorical values encountered during processing are

Variables	Levels		
CHOICE (3 levels)	1.000	2.000	3.000

Discrete Choice Analysis

```
Dependent Variable : CHOICE
Input Records      : 29
Records for Analysis : 29
```

**Sample Split**

Category Choices	
1	15
2	6
3 (REFERENCE)	8
Total	29

Log-Likelihood Iteration History

```

Log-Likelihood at Iteration1 : -31.860
Log-Likelihood at Iteration2 : -25.808
Log-Likelihood at Iteration3 : -25.779
Log-Likelihood at Iteration4 : -25.779
Log-Likelihood at Iteration5 : -25.779
Log-Likelihood                : -25.779

```

**Information Criteria**

```

AIC      : 59.557
Schwarz's BIC : 65.026

```

**Parameter Estimates**

Parameter	Estimate	Standard Error	Z	p-value
1 TIME	-0.012	0.020	-0.575	0.565
2 COST	-0.567	0.222	-2.550	0.011
3 CONSTANT	1.510	0.608	2.482	0.013
3 CONSTANT	-0.865	0.675	-1.282	0.200

**Odds Ratio Estimates**

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval Lower	Upper
1 TIME	0.988	0.020	0.950	1.029
2 COST	0.567	0.126	0.367	0.877

```

Log-Likelihood of Constants only Model = LL(0) : -29.645
2*[LL(N)-LL(0)] : 7.732
df : 2
p-value : 0.021

```

```

McFadden's Rho-squared : 0.130
Cox and Snell R-square : 0.234
Naglekerke's R-square : 0.269

```

**Wald Tests on Effects Across all Choices**

Effect	Wald Statistic	Chi-square Significance	df
3 CONSTANT	8.630	0.013	2.000

**Covariance Matrix**

	1	2	3	4
1	0.000			
2	0.001	0.049		
3	-0.001	-0.082	0.370	
4	-0.005	0.056	0.046	0.455

**Correlation Matrix**

	1	2	3	4
1	1.000			
2	0.130	1.000		
3	-0.053	-0.606	1.000	
4	-0.350	0.372	0.113	1.000



### Example 9

#### By-Choice Data Format

In the standard data layout, there is one data record per case that contains information on every alternative open to a chooser. With a large number of alternatives, this can quickly lead to an excessive number of variables. A convenient alternative is to organize data by choice; with this data layout, there is one record per alternative and as many as NCAT records per case. The data set *CHOICE2* organizes the *CHOICE* data of the Discrete Choice Models example in this way. If you analyze the differences between the two data sets, you will see that they are similar to those between the split-plot and multivariate layout for the repeated measures design (see Statistics II, Chapter 3, Linear Models II - Analysis of Variance). To set up the same problem in a by-choice layout, input the following:

```
USE CHOICE2
LOGIT
NCAT 3
ALT NCHOICES
MODEL CHOICE=TIME+COST ;
ESTIMATE
```

The by-choice format requires that the dependent variable appear with the same value on each record pertaining to the case. An ALT variable (here *NCHOICES*) indicating the number of records for this case must also appear on each record. The by-choice organization results in fewer variables on the data set, with the savings increasing with the number of alternatives. However, there is some redundancy in that certain data values are repeated on each record. The best reason for using a by-choice format is to handle varying numbers of alternatives per case. In this situation, there is no need to shuffle data values or to be concerned with choice order.

With the by-choice data format, the NCAT statement is required; it is the only way for LOGIT to know the number of alternatives to expect per case. For varying numbers of alternatives per case, the ALT statement is also required, although we use it here with the same number of alternatives.

```
USE CHOICE2
LOGIT
CATEGORY SEX$
NCAT 3
ALT NCHOICES
MODEL CHOICE=TIME+COST ; AGE+SEX$
ESTIMATE
```

Because the number of alternatives (ALT) is the same for each case in this example, the output is the same as the "Mixed Parameters" example.

### *Weighting Choice-Based Samples*

For estimation of the slope coefficients of the discrete choice model, weighting is not required even in choice-based samples. For predictive purposes, however, weighting is necessary to forecast aggregate shares, and it is also necessary for consistent estimation of the alternative specific dummies (Manski and Lerman, 1977).

The appropriate weighting procedure for choice-based sample logit estimation requires that the sum of the weights equal the actual number of observations retained in the estimation sample. For choice-based samples, the weight for any observation choosing the  $i$ th option is  $W_i = S_i/s_i$ , where  $S_i$  is the population share choosing the  $i$ th option and  $s_i$  is the choice-based sample share choosing the  $i$ th option.

As an example, suppose theatergoers make up 10% of the population and we have a choice-based sample consisting of 100 theatergoers ( $Y = 1$ ) and 100 non-theatergoers ( $Y = 0$ ). Although theatergoers make up only 10% of the population, they are heavily oversampled and make up 50% of the study sample. Using the above formulas, the correct weights would be

$$W_0 = 0.9/0.5 = 1.8$$

$$W_1 = 0.1/0.5 = 0.2$$

and the sum of the weights would be  $100 * 1.8 + 100 * 0.2 = 200$ , as required. To handle such samples, LOGIT permits non-integer weights and does not truncate them to integers.

### *Example 10* *Stepwise Regression*

LOGIT offers forward and backward stepwise logistic regression with single stepping as an option. The simplest way to initiate stepwise regression is to substitute START for ESTIMATE following a MODEL statement and then proceed with stepping with the STEP command, just as in GLM or Regression.

An upward step consists of three components. First, the current model is estimated to convergence. The procedure is exactly the same as regular estimation. Second, score statistics for each additional effect are conducted, adjusted for variables already in the

model. The joint significance of all additional effects together is also computed. Finally, the effect with the smallest significance level for its score statistic is identified. If this significance level is below the ENTER option (0.05 by default), the effect is added to the model.

A downward step also consists of three computational segments. First, the model is estimated to convergence. Then Wald statistics are computed for each effect in the model. Finally, the effect with the largest  $p$  value for its Wald test statistic is identified. If this significance level is above the REMOVE criterion (by default 0.10), the effect is removed from the model.

If you require certain effects to remain in the model regardless of the outcome of the Wald test, force them into the model by listing them first in the model and using the FORCE option of START. It is important to set the ENTER and REMOVE criteria carefully because it is possible to have a variable cycle in and out of a model repeatedly. Each step of the analysis consists of AIC, AIC (corrected), Schwarz's BIC values which are tools for model selection. The defaults are:

```
START / ENTER = .05, REMOVE = .10
```

although Hosmer and Lemeshow use

```
START / ENTER = .15, REMOVE = .20
```

in the example we reproduce below.

Hosmer and Lemeshow use stepwise regression in their search for a model of low birth weight discussed in the "Binary Logit" section. We conduct a similar analysis.

The input is:

```
USE HOSLEM
LOGIT
CATEGORY RACE
MODEL LOW=CONSTANT+PTL+LWT+HT+RACE+SMOKE+UI+AGE+FTV
START / ENTER=.15, REMOVE=.20
STEP / AUTO
STOP
```

## The output is:

## Logistic Regression

## Stepwise Selection of Variables

## Stepping Parameters

Significance to Include : 0.150  
 Significance to Remove : 0.200  
 Number of Effects to Force : 1  
 Maximum Number of Steps : 10

## Direction : Both

Categorical values encountered during processing are

Variables	Levels
PACE (3 levels)	1.000 2.000 3.000
LOW (2 levels)	0.000 1.000

Categorical variables are effects coded with the highest value as reference. Effects coding is in force for the categorical independent variables in your model. Parameters and odds ratios are easier to interpret for dummy coded categorical. Unless you have specific reasons for requesting effects coding, we suggest that you re-issue the category statement with the /dummy option and re-fit your model. See Hosmer & Lemeshow, for more information.

## Binary Stepwise LOGIT Analysis

Dependent Variable : LOW  
 Input Records : 189  
 Records for Analysis : 189

## Sample Split

Category Choices	
0 (REFERENCE)	130
1 (RESPONSE)	59
Total	189

## Step 0

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1 : -131.005  
 Log-Likelihood at Iteration2 : -117.366  
 Log-Likelihood at Iteration3 : -117.336  
 Log-Likelihood at Iteration4 : -117.336  
 Log-Likelihood : -117.336

## Information Criteria

AIC : 236.672  
 Schwarz's BIC : 239.914

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval
					Lower Upper
1 CONSTANT	-0.790	0.157	-5.033	0.000	-1.098 -0.482



## Score Tests on Effects not in Model

Effect	Score Statistic	Chi-square Significance	df
2 PTL	7.267	0.007	1.000
3 LMT	5.438	0.020	1.000
4 HT	4.388	0.036	1.000
5 RACE	5.005	0.082	2.000
6 SMOKE	4.924	0.026	1.000
7 UI	5.401	0.020	1.000
8 AGE	2.674	0.102	1.000
9 FTV	0.749	0.387	1.000
Joint Score	30.959	0.000	9.000

## Step 1

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-131.005
Log-Likelihood at Iteration2	-114.024
Log-Likelihood at Iteration3	-113.946
Log-Likelihood at Iteration4	-113.946
Log-Likelihood	-113.946

## Information Criteria

AIC	231.893
Schwarz's BIC	238.376

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	-0.964	0.175	-5.511	0.000	-1.307	-0.621
2 PTL	0.802	0.317	2.528	0.011	0.180	1.423

## Score Tests on Effects not in Model

Effect	Score Statistic	Chi-square Significance	df
3 LMT	4.113	0.043	1.000
4 HT	4.722	0.030	1.000
5 RACE	5.359	0.069	2.000
6 SMOKE	3.164	0.075	1.000
7 UI	3.161	0.075	1.000
8 AGE	3.478	0.062	1.000
9 FTV	0.577	0.448	1.000
Joint Score	24.772	0.002	8.000

## Step 2

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-131.005
Log-Likelihood at Iteration2	-111.911
Log-Likelihood at Iteration3	-111.792
Log-Likelihood at Iteration4	-111.792
Log-Likelihood	-111.792

## Information Criteria

AIC	229.583
Schwarz's BIC	239.309

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	-1.062	0.184	-5.764	0.000	-1.423	-0.701
2 PTL	0.823	0.318	2.585	0.010	0.199	1.447
3 HT	1.272	0.616	2.066	0.039	0.066	2.479

## Score Tests on Effects not in Model

Effect	Score Statistic	Chi-square	
		Significance	df
4 LWT	6.900	0.009	1.000
5 RACE	4.882	0.087	2.000
6 SMOKE	3.117	0.078	1.000
7 UI	4.225	0.040	1.000
8 AGE	3.448	0.063	1.000
9 FTV	0.370	0.543	1.000
Joint Score	20.658	0.004	7.000

Step 3

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-131.005
Log-Likelihood at Iteration2	-108.523
Log-Likelihood at Iteration3	-107.987
Log-Likelihood at Iteration4	-107.982
Log-Likelihood at Iteration5	-107.982
Log-Likelihood	-107.982

## Information Criteria

AIC	223.964
Schwarz's BIC	236.931

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	1.093	0.841	1.299	0.194	-0.556	2.742
2 PTL	0.726	0.328	2.213	0.027	0.083	1.368
3 HT	1.856	0.705	2.633	0.008	0.474	3.238
4 LWT	-0.017	0.007	-2.560	0.010	-0.030	-0.004

## Score Tests on Effects not in Model

Effect	Score Statistic	Chi-square	
		Significance	df
5 RACE	5.266	0.072	2.000
6 SMOKE	2.857	0.091	1.000
7 UI	3.081	0.079	1.000
8 AGE	1.895	0.169	1.000
9 FTV	0.118	0.732	1.000
Joint Score	14.395	0.026	6.000

Step 4

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1	-131.005
Log-Likelihood at Iteration2	-106.169
Log-Likelihood at Iteration3	-105.434
Log-Likelihood at Iteration4	-105.425
Log-Likelihood at Iteration5	-105.425
Log-Likelihood	-105.425

## Information Criteria

AIC | 222.850  
 Schwarz's BIC | 242.301

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	1.405	0.900	1.560	0.119	-0.360	3.170
2 PTL	0.746	0.328	2.278	0.023	0.104	1.389
3 HT	1.805	0.714	2.530	0.011	0.407	3.204
4 LWT	-0.018	0.007	-2.607	0.009	-0.032	-0.004
5 RACE_1	-0.518	0.237	-2.190	0.029	-0.983	-0.054
6 RACE_2	0.569	0.318	1.787	0.074	-0.055	1.193

## Score Tests on Effects not in Model

Effect	Score Statistic	Chi-square Significance	df
6 SMOKE	5.936	0.015	1.000
7 UI	3.265	0.071	1.000
8 AGE	1.019	0.313	1.000
9 FTV	0.025	0.873	1.000
Joint Score	9.505	0.050	4.000

Step 5

## Log-Likelihood Iteration History

Log-Likelihood at Iteration1 | -131.005  
 Log-Likelihood at Iteration2 | -103.581  
 Log-Likelihood at Iteration3 | -102.468  
 Log-Likelihood at Iteration4 | -102.449  
 Log-Likelihood at Iteration5 | -102.449  
 Log-Likelihood | -102.449

## Information Criteria

AIC | 218.898  
 Schwarz's BIC | 241.590

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	0.851	0.913	0.933	0.351	-0.938	2.641
2 PTL	0.602	0.335	1.797	0.072	-0.055	1.260
3 HT	1.745	0.695	2.511	0.012	0.383	3.107
4 LWT	-0.017	0.007	-2.418	0.016	-0.030	-0.003
5 RACE_1	-0.734	0.263	-2.790	0.005	-1.249	-0.218
6 RACE_2	0.557	0.324	1.720	0.085	-0.078	1.191
7 SMOKE	0.946	0.395	2.396	0.017	0.172	1.720

## Score Tests on Effects not in Model

Effect	Score Statistic	Chi-square Significance	df
7 UI	3.034	0.082	1.000
8 AGE	0.781	0.377	1.000
9 FTV	0.014	0.904	1.000
Joint Score	3.711	0.294	3.000

## Step 6

## Log-Likelihood Iteration History

```

Log-Likelihood at Iteration1  -131.005
Log-Likelihood at Iteration2  -102.280
Log-Likelihood at Iteration3  -101.017
Log-Likelihood at Iteration4  -100.993
Log-Likelihood at Iteration5  -100.993
Log-Likelihood at Iteration6  -100.993
Log-Likelihood                 -100.993

```

## Information Criteria

```

AIC      | 217.986
Schwarz's BIC | 243.920

```

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	0.654	0.921	0.710	0.477	-1.151	2.460
2 PTL	0.503	0.341	1.475	0.140	-0.166	1.172
3 HT	1.855	0.695	2.669	0.008	0.493	3.217
4 LWT	-0.016	0.007	-2.320	0.020	-0.029	-0.002
5 RACE_1	-0.741	0.265	-2.797	0.005	-1.260	-0.222
6 RACE_2	0.585	0.323	1.811	0.070	-0.048	1.218
7 SMOKE	0.939	0.399	2.354	0.019	0.157	1.720
8 UI	0.786	0.456	1.721	0.085	-0.109	1.680

## Score Tests on Effects not in Model

Effect	Score Statistic	Chi-square Significance	df
8 AGE	0.553	0.457	1.000
9 FTV	0.056	0.813	1.000
Joint Score	0.696	0.706	2.000

## Final Model Summary

## Log-Likelihood Iteration History

```

Log-Likelihood at Iteration1  -131.005
Log-Likelihood at Iteration2  -102.280
Log-Likelihood at Iteration3  -101.017
Log-Likelihood at Iteration4  -100.993
Log-Likelihood at Iteration5  -100.993
Log-Likelihood at Iteration6  -100.993
Log-Likelihood                 -100.993

```

## Information Criteria

```

AIC      | 217.986
Schwarz's BIC | 243.920

```

## Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value	95 % Confidence Interval	
					Lower	Upper
1 CONSTANT	0.654	0.921	0.710	0.477	-1.151	2.460
2 PTL	0.503	0.341	1.475	0.140	-0.166	1.172
3 HT	1.855	0.695	2.669	0.008	0.493	3.217
4 LWT	-0.016	0.007	-2.320	0.020	-0.029	-0.002
5 RACE_1	-0.741	0.265	-2.797	0.005	-1.260	-0.222
6 RACE_2	0.585	0.323	1.811	0.070	-0.048	1.218
7 SMOKE	0.939	0.399	2.354	0.019	0.157	1.720
8 UI	0.786	0.456	1.721	0.085	-0.109	1.680



## Odds Ratio Estimates

Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
			Lower	Upper
2 PTL	1.654	0.564	0.847	3.229
3 HT	6.392	4.443	1.637	24.964
4 LWT	0.984	0.007	0.971	0.998
5 RACE_1	0.477	0.126	0.284	0.801
6 RACE_2	1.795	0.579	0.953	3.379
7 SMOKE	2.557	1.019	1.170	5.586
8 UI	2.194	1.001	0.897	5.367

Log-Likelihood of Constants only Model = LL(0) : -117.336  
 2\*[LL(N)-LL(0)] : 32.686  
 df : 7  
 p-value : 0.000

McFadden's Rho-squared : 0.139  
 Cox and Snell R-square : 0.159  
 Naglekerke's R-square : 0.223

Not all logistic regression programs compute the variable addition statistics in the same way, so minor differences in output are possible. Our results listed in the *Chi-Square Significance* column of the first step, for example, correspond to H&L's first row in their Table 4.15; the two sets of results are very similar but not identical. While our method yields the same final model as H&L, the order in which variables are entered is not the same because intermediate *p* values differ slightly. Once a final model is arrived at, it is re-estimated to give true maximum likelihood estimates.

## Example 11

### Hypothesis Testing

Two types of hypothesis tests are easily conducted in LOGIT: the likelihood ratio (LR) test and the Wald test. The tests are discussed in numerous statistics books, sometimes under varying names. Accounts can be found in Maddala's text (2001), Cox and Hinkley (1979), Rao (1973), Engel (1984), and Breslow and Day (1980). Here we provide some elementary examples.

### Likelihood-Ratio Test

The likelihood-ratio test is conducted by fitting two nested models (the restricted and the unrestricted) and comparing the log-likelihoods at convergence. Typically, the unrestricted model contains a proposed set of variables, and the restricted model omits a selected subset, although other restrictions are possible. The test statistic is twice the difference of the log-likelihoods and is chi-squared with degrees of freedom equal to

the number of restrictions imposed. When the restrictions consist of excluding variables, the degrees of freedom are equal to the number of parameters set to 0.

If a model contains a constant, LOGIT automatically calculates a likelihood-ratio test of the null hypothesis that all coefficients except the constant are 0. It appears on a line that looks like:

```
2*[LL(N)-LL(0)] = 26.586 with 5 df, Chi-sq p-value = 0.00007
```

This example line states that twice the difference between the likelihood of the estimated model and the “constants only” model is 26.586, which is a chi-squared deviate on five degrees of freedom. The  $p$  value indicates that the null hypothesis would be rejected.

To illustrate use of the LR test, consider a model estimated on the low birth weight data (see the “Binary Logit” example). Assuming CATEGORY=RACE, compare the following model

```
MODEL LOW CONSTANT + LWD + AGE + RACE + PTD
```

with

```
MODEL LOW CONSTANT + LWD + AGE
```

The null hypothesis is that the categorical variable *RACE*, which contributes two parameters to the model, and *PTD* are jointly 0. The model likelihoods are -104.043 and -112.143, and twice the difference (16.20) is chi-squared with three degrees of freedom under the null hypothesis. This value can also be more conveniently calculated by taking the difference of the LR test statistics reported below the parameter estimates and the difference in the degrees of freedom. The unrestricted model above has  $G = 26.587$  with five degrees of freedom, and the restricted model has  $G = 10.385$  with two degrees of freedom. The difference between the  $G$  values is 16.20, and the difference between degrees of freedom is 3.

Although LOGIT will not automatically calculate  $LR$  statistics across separate models, the  $p$  value of the result can be obtained with the command:

```
CALC 1-XCF(16.2,3)
```

### Wald Test

The Wald test is the best known inferential procedure in applied statistics. To conduct a Wald test, we first estimate a model and then pose a linear constraint on the parameters estimated. The statistic is based on the constraint and the appropriate

elements of the covariance matrix of the parameter vector. A test of whether a single parameter is 0 is conducted as a Wald test by dividing the squared coefficient by its variance and referring the result to a chi-squared distribution on one degree of freedom. Thus, each  $z$  ratio is itself the square root of a simple Wald test. Following is an example:

```
USE HOSLEM
LOGIT
CATEGORY RACE
MODEL LOW=CONSTANT+LWD+AGE+RACE+PTD
ESTIMATE
HYPOTHESIS
CONSTRAIN PTD=0
CONSTRAIN RACE[1]=0
CONSTRAIN RACE[2]=0
TEST
```

The output is (minus the estimation stage):

Hypothesis Tests  
Entering Hypothesis Procedure

#### Linear Restriction System

EQN	Parameter				
	1	2	3	4	5
1	0.000	0.000	0.000	0.000	0.000
2	0.000	0.000	0.000	1.000	0.000
3	0.000	0.000	0.000	0.000	1.000

#### Linear Restriction System

EQN	6	Parameter RHS	Q
1	1.000	0.000	1.515
2	0.000	0.000	-0.442
3	0.000	0.000	0.464

#### General Linear Wald Test Results

Chi-square Statistic : 15.104  
df : 3  
p-value : 0.002

Note that this statistic of 15.104 is close to the  $LR$  statistic of 16.2 obtained for the same hypothesis in the previous section. Although there are three separate **CONSTRAIN** lines in the **HYPOTHESIS** paragraph above, they are tested jointly in a single test. To test each restriction individually, place a **TEST** after each **CONSTRAIN**. The restrictions being tested are each entered with separate **CONSTRAIN** commands. These can include any linear algebraic expression without parentheses involving the parameters. If interactions were present on the **MODEL** statement, they can also appear on the **CONSTRAIN** statement. To reference dummies generated from categorical covariates,



use square brackets, as in the example for *RACE*. This constraint refers to the coefficient labeled *RACE-1* in the output.

More elaborate tests can be posed in this framework. For example,

```
CONSTRAIN 7*LWD - 4.3*AGE + 1.5*RACE[2] = -5
```

or

```
CONSTRAIN AGE + LWD = 1
```

For multinomial models, the architecture is a little different. To reference a variable that appears in more than one parameter vector, it is followed with curly braces around the number corresponding to the *Choice Group*. For example,

```
CONSTRAIN CONSTANT{1} - CONSTANT{2} = 0
CONSTRAIN AGE{1} - AGE{2} = 0
```

### Comparisons between Tests

The Wald and likelihood-ratio tests are classical testing methods in statistics. The properties of the tests are based on asymptotic theory, and in the limit, as sample sizes tend to infinity, the tests give identical results. In small samples, there will be differences between results and conclusions, as has been emphasized by Hauck and Donner (1977). Given a choice, which test should be used?

Most statisticians favor the LR test over the Wald for three reasons. First, the likelihood is the fundamental measure on which model fitting is based. Cox and Oakes (1984) illustrate this preference when they use the likelihood profile to determine confidence intervals for a parameter in a survival model. Second, Monte Carlo studies suggest that the LR test is more reliable in small samples. Finally, a nonlinear constraint can be imposed on the parameter estimates and simply tested by estimating restricted and unrestricted models. See the "Quantiles" example for an illustration involving LD50 values. Also, you can use the FUNPAR option in NONLIN to do the same thing.

Why bother with the Wald test, then? One reason is simplicity and computational cost. The LR test requires estimation of two models to final convergence for a single test, and each additional test requires another full estimation. By contrast, any number of Wald tests can be run on the basis of one estimated model, and they do not require an additional pass through the data.



### Example 12

#### Tackling different data format in Logistic Regression

So far, we have come across the data format in which each case (row) corresponds to a single trial and the response in that trial is indicated by a variable (binary or p-array).

Case no	Response	Explanatory
1	a	(x11 x21...xp1)
2	b	(x12 x22...xp2)
.	a	
.	b	
.	.	
N	a	(x1N x2N...xpN)

Now, if the dependent variable specifies two variables: number of events and number of observations; in other words if the event is binomially distributed with the number of trials given by the number of observations, what should be the correct syntax for handling such data in SYSTAT?

Clearly the second kind of data format is as follows:

Case no	Trial	Event	Explanatory
1	n1	s1	(x11 x21...xp1)
2	n2	s2	(x12 x22...xp2)
.	.	.	.
.	.	.	.
N	nN	sN	(x1N x2N...xpN)

A possible solution to the query is creation of an appropriate data file in SYSTAT. To do this a suitable example is given below with necessary explanations. The *TARGET* data set is hypothetical. It describes the success of an arrow throwing machine to hit the target. The aim is to analyze the relationship between the probability of success of the machine and the height at which the machine is placed (in centimeters), and the force applied (in newtons).

In *TARGET* there is no response variable available explicitly and so it cannot be readily handled in SYSTAT. But just by adding one more variable, the analysis can be done in SYSTAT. The data modification is independent of the number of explanatory variables.

The input is:

```
USE TARGET
LET eventtype = 1
ESAVE target1.syz
USE target.syz
LET eventtype = 0
LET noofevents = nooftrails-noofevents
ESAVE target2.syz
APPEND target1 target2
ESAVE app.syz
```

The resultant data *APP* contains a response variable *EVENTTYPE*. Now each case corresponds to experiment number and events frequency with event type (0 or 1). A few data points from *APP* are as follows:

EXPTNO	NOOFTRAILS	NOOFEVENTS	HEIGHT	FORCE	EVENTTYPE
1.000	50.000	35.000	136.315	0.577	1.000
2.000	50.000	33.000	138.026	0.622	1.000
3.000	50.000	35.000	137.820	0.501	1.000
.	.	.	.	.	.
.	.	.	.	.	.
48.000	50.000	32.000	139.202	0.745	1.000
49.000	50.000	32.000	135.484	0.708	1.000
50.000	50.000	31.000	137.693	0.746	1.000
1.000	50.000	15.000	136.315	0.577	0.000
2.000	50.000	17.000	138.026	0.622	0.000
3.000	50.000	15.000	137.820	0.501	0.000
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
48.000	50.000	18.000	139.202	0.745	0.000
49.000	50.000	18.000	135.484	0.708	0.000
50.000	50.000	19.000	137.693	0.746	0.000

Now to analyze APP the input is:

```
USE APP
FREQ NOOFEVENTS
LOGIT
MODEL eventtype = constant + height + force
ESTIMATE
```

Computation

Algorithms

## The output is:

### Logistic Regression

Case frequencies determined by value of variable NOOFEVENTS

Categorical values encountered during processing are

Variables	Levels
-----	-----
EVENTTYPE (2 levels)	0.000 1.000

### Binary LOGIT Analysis

Dependent Variable : EVENTTYPE  
 Analysis is Weighted by : NOOFEVENTS  
 Sum of Weights : 2500.000  
 Input Records : 100  
 Records for Analysis : 100

### Sample Split

Category	Count	Weighted Count
-----	-----	-----
0 (REFERENCE)	50	1626
1 (RESPONSE)	50	874
Total	100	2500.000

### Log-Likelihood Iteration History

Log-Likelihood at Iteration1 : -1732.868  
 Log-Likelihood at Iteration2 : -1618.041  
 Log-Likelihood at Iteration3 : -1617.935  
 Log-Likelihood at Iteration4 : -1617.935  
 Log-Likelihood : -1617.935

### Information Criteria

AIC : 3241.871  
 Schwarz's BIC : 3249.686

### Parameter Estimates

Parameter	Estimate	Standard Error	Z	p-value
-----	-----	-----	-----	-----
1 CONSTANT	1.840	3.912	0.470	0.638
2 HEIGHT	-0.008	0.028	-0.296	0.767
3 FORCE	-0.110	0.568	-0.195	0.846

### Odds Ratio Estimates

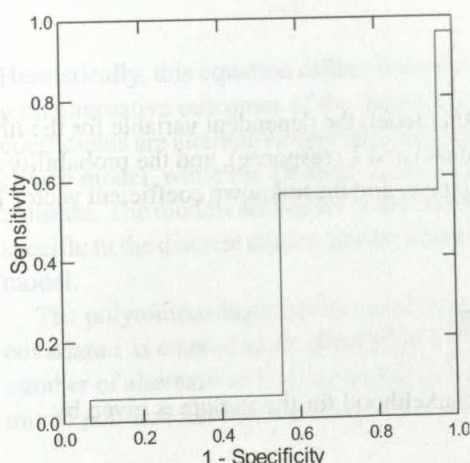
Parameter	Odds Ratio	Standard Error	95 % Confidence Interval	
-----	-----	-----	Lower	Upper
2 HEIGHT	0.992	0.028	0.938	1.048
3 FORCE	0.895	0.508	0.294	2.724

Log-Likelihood of Constants only Model = LL(0) : -1617.997  
 2\*[LL(N)-LL(0)] : 0.123  
 df : 2  
 p-value : 0.941

McFadden's Rho-squared : 0.000  
 Cox and Snell R-square : 0.001  
 Naglekerke's R-square : 0.001



Receiver Operating Characteristic Curve



Area under ROC Curve : 0.502

## Computation

### Algorithms

LOGIT uses Gauss Newton methods for maximizing the likelihood. By default, two tolerance criteria must be satisfied: the maximum value for relative coefficient changes must fall below 0.001, and the Euclidean norm of the relative parameter change vector must also fall below 0.001. By default, LOGIT uses the second derivative matrix to update the parameter vector. In discrete choice models, it may be preferable to use a first derivative approximation to the Hessian instead. This option, popularized by Berndt, Hall, Hall, and Hausman (1974), will be noted if it is used by the program. BHHH uses the summed outer products of the gradient vector in place of the Hessian matrix and generally will converge much more slowly than the default method.

## Missing Data

Cases with missing data on any variables included in a model are deleted.

### Basic Formulas

For the binary logistic regression model, the dependent variable for the  $i$ th case is  $Y_i$ , taking on values of 0 (nonresponse) and 1 (response), and the probability of response is a function of the covariate vector  $x_i$  and the unknown coefficient vector  $\beta$ . We write this probability as:

$$Prob(Y_i = 1 | x_i) = \frac{e^{x_i \beta}}{1 + e^{x_i \beta}}$$

and abbreviate it as  $P_i$ . The log-likelihood for the sample is given by

$$LL(\beta) = \sum_{i=1}^n Y_i \log P_i + (1 - Y_i) \log (1 - P_i)$$

For the polytomous multinomial logit, the integer-valued dependent variable ranges from 1 to  $k$ , and the probability that the  $i$ th case has  $Y = m$ , where  $1 \leq m \leq k$  is:

$$Prob(Y_i = m | x_i) = \frac{e^{x_i \beta_m}}{\sum_{j=1}^k e^{x_i \beta_j}}$$

In this model,  $k$  is fixed for all cases, there is a single covariate vector  $x_i$ , and  $k$   $\beta_j$  parameter vectors are estimated. This last equation is identified by normalizing  $\beta_k$  to 0.

McFadden's discrete choice model represents a distinct variant of the logit model based on Luce's (1959) probabilistic choice model. Each subject is observed to make a choice from a set  $C_i$  consisting of  $J_i$  elements. Each element is characterized by a separate covariate vector of attributes  $Z_k$ . The dependent variable  $Y_i$  ranges from 1 to  $J_i$ , with  $J_i$  possibly varying across subjects, and the probability that  $Y_i = k$ , where  $1 \leq k \leq J_i$  is a function of the attribute vectors  $Z_1, Z_2, \dots, Z_{J_i}$  and the parameter vector  $\beta$ . The probability that the  $i$ th subject chooses element  $m$  from his choice set is:

$$Prob(Y_i = m|Z) = \frac{e^{Z_m\beta}}{\sum_{j \in C_i} e^{Z_j\beta}}$$

Heuristically, this equation differs from the previous one in the components that vary with alternative outcomes of the dependent variable. In the polytomous logit, the coefficients are alternative-specific and the covariate vector is constant; in the discrete choice model, while the attribute vector is alternative-specific, the coefficients are constant. The models also differ in that the range of the dependent variable can be case-specific in the discrete choice model, while it is constant for all cases in the polytomous model.

The polytomous logit can be recast as a discrete choice model in which each covariate  $x$  is entered as an interaction with an alternative-specific dummy, and the number of alternatives is constant for all cases. This reparameterization is used for the mixed polytomous discrete choice model.

### Regression Diagnostics Formulas

The SAVE command issued before the deciles of risk command (DC) produces a SYSTAT save file with a number of diagnostic quantities computed for each case in the input data set. Computations are always conducted on the assumption that each covariate pattern is unique. The following formulas are based on the binary dependent variable  $y_i$ , which is either 0 or 1, and fitted probabilities  $P_i$ , obtained from the basic logistic equation.

LEVERAGE(1) is the diagonal element of Pregibon's (1981) *hat* matrix, with formulas given by Hosmer and Lemeshow (2000) as their equations (5.12) and (5.13). It is defined as  $b_j v_j$ , where

$$b_j = x_j'(X'VX)^{-1}x_j$$

and  $x_j$  is the covariate vector for the  $x$ th case,  $\mathbf{X}$  is the data matrix for the sample including a constant, and  $\mathbf{V}$  is a diagonal matrix with general  $\mathbf{A}$  element  $P_i(1 - P_i)$ , the fitted probability for the  $i$ th case.  $b_j$  is our LEVERAGE(2).

$$v_j = \hat{P}_i(1 - \hat{P}_i)$$

Thus LEVERAGE(L) is given by

$$h_j = v_j b_j$$

The PEARSON residual is

$$r_j = \frac{y_i - \hat{p}_i}{\sqrt{p_i(1 - p_i)}}$$

The VARIANCE of the residual is

$$v_j(1 - h_j)$$

and the standardized residual STANDARD is

$$r_{sj} = \frac{r_j}{\sqrt{1 - h_j}}$$

The DEVIANCE residual is defined as

$$d_j = \sqrt{2|\ln(p_j)|}$$

for  $y_j = 1$  and

$$d_j = -2\sqrt{|\ln(1 - p_j)|}$$

otherwise.

DELDSTAT is the change in deviance and is

$$\nabla D_j = d_j^2 / (1 - h_j)$$

DELPSTAT is the change in Pearson chi-square:

$$\nabla \chi^2 = r_{sj}^2$$

is a measure proposed by Pregibon, and



The final three saved quantities are measures of the overall change in the estimated parameter vector  $\beta$ .

$$DELBETA(1) = r_{sj}^2 h_j / (1 - h_j)$$

## References

- Agresti, A. (2002). *Categorical data analysis*. 2nd ed. New York: John Wiley & Sons.
- \*Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika*, 71, 1–10.
- Amemiya, T. (1981). Qualitative response models: A survey. *Journal of Economic Literature*, 1483–1536.
- Anderson, J.A. (1972). Separate sample logistic discrimination: *Biometrika*, 59(1): 19-35.
- Begg, C. and Gray, R. (1984). Calculation of polychotomous logistic regression parameters using individualized regressions. *Biometrika*, 71, 11–18.
- Beggs, S., Cardell, N. S., and Hausman, J. A. (1981). Assessing the potential demand for electric cars. *Journal of Econometrics*, 16, 1–19.
- Ben-Akiva, M. and Lerman, S. (1985). *Discrete choice analysis*. Cambridge, Mass.: MIT Press.
- Berndt, E. K., Hall, B. K., Hall, R. E., and Hausman, J. A. (1974). Estimation and inference in non-linear structural models. *Annals of Economic and Social Measurement*, 3, 653–665.
- Breslow, N. (1982). Covariance adjustment of relative-risk estimates in matched studies. *Biometrics*, 38, 661–672.
- Breslow, N. and Day, N. E. (1980). *Statistical methods in cancer research, vol. II: The design and analysis of cohort studies*. Lyon: IARC.
- Breslow, N., Day, N. E., Halvorsen, K.T, Prentice, R.L., and Sabai, C. (1978). Estimation of multiple relative risk functions in matched case-control studies. *American Journal of Epidemiology*, 108, 299–307.
- Burnham, K.P., and Anderson, D.R. (1992). *Data-based selection of an appropriate biological model: the key to modern data analysis*. Pages 16-30 in Wildlife 2001: Populations, McCullough, D.R. and R.H. Barrett (eds.). Elsevier Science Publishers, Ltd. London, England.
- \*Burnham, K.P., and Anderson, D.R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer-Verlag.

- Carson, R., Hanemann, M. and Steinberg, S. (1990). A discrete choice contingent valuation estimate of the value of kenai king salmon. *Journal of Behavioral Economics*, 19, 53–68.
- Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, 47, 225–238.
- Cook, D. R. and Weisberg, S. (1982). *Residuals and influence in regression*. New York: Chapman and Hall.
- Coslett, S. R. (1980). Efficient estimation of discrete choice models. In C. Manski and D. McFadden, Eds., *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass.: MIT Press.
- Cox, D. R. (1975). Partial likelihood. *Biometrika*, 62, 269–276.
- Cox, D. R. and Hinkley, D.V. (1979). *Theoretical statistics*. London: Chapman and Hall.
- Cox, D. R. and Oakes, D. (1984). *Analysis of survival data*. New York: Chapman and Hall.
- Cox, D. R. and Snell, E.J. (1989). *The analysis of binary data*. 2nd ed. Methuen, London.
- Domencich, T. and McFadden, D. (1975). *Urban travel demand: A behavioral analysis*. Amsterdam: North-Holland.
- Engel, R. F. (1984). Wald, likelihood ratio and Lagrange multiplier tests in econometrics. In Z. Griliches and M. Intriligator, Eds., *Handbook of Econometrics*. New York: North-Holland.
- Finney, D. J. (1978). *Statistical method in biological assay*. London: Charles Griffin.
- \*Hanley, J. A. and McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143, 29–36.
- \*Hauck, W. W. (1980). A note on confidence bands for the logistic response curve. *American Statistician*, 37, 158–160.
- Hauck, W. W. and Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *Journal of the American Statistical Association*, 72, 851–853.
- Hensher, D. and Johnson, L. W. (1981). *Applied discrete choice modelling*. London: Croom Helm.
- Hoffman, S. and Duncan, G. (1988). Multinomial and conditional logit discrete choice models in demography. *Demography*, 25, 415–428.
- Hosmer, D. W. and Lemeshow, S. (2000). *Applied logistic regression*, 2nd ed. New York: John Wiley & Sons.
- Hubert, J. J. (1984). *Bioassay*, 2nd ed. Dubuque, Iowa: Kendall-Hunt.
- \*Kalbfleisch, J. D. and Prentice, R. L. (2002). *The statistical analysis of failure time data*. New York: John Wiley & Sons.
- Kleinbaum, D., Kupper, L., and Chambliss, L. (1982). Logistic regression analysis of epidemiologic data: Theory and practice. *Communications in Statistics: Theory and Methods*, 11, 485–547.



- Luce, R. D. (2005). *Individual choice behavior: A theoretical analysis*. Mineola, N.Y.: Dover Press.
- Luft, H., Garnick, D., Peltzman, D., Phibbs, C., Lichtenberg, E., and McPhee, S. (1988). *The sensitivity of conditional choice models for hospital care to estimation technique*. Draft, Institute for Health Policy Studies. San Francisco: University of California.
- Maddala, G. S. S. (1986). *Limited-dependent and qualitative variables in econometrics*. Cambridge University Press.
- Maddala, G. S. S. (2001). *Introduction to econometrics*. New York: John Wiley & Sons.
- Manski, C. and Lerman, S. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, 8, 1977–1988.
- Manski, C. and McFadden, D. (1980). Alternative estimators and sample designs for discrete choice analysis. In C. Manski and D. McFadden (eds.), *Structural Analysis of Discrete Data with Econometric Applications*. Cambridge, Mass.: MIT Press.
- \*Manski, C. and McFadden, D., eds. (1981). *Structural analysis of discrete data with econometric applications*. Cambridge, Mass.: MIT Press.
- McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (ed.), *Frontiers in Econometrics*. New York: Academic Press.
- McFadden, D. (1976). Quantal choice analysis: A survey. *Annals of Economic and Social Measurement*, 5, 363–390.
- McFadden, D. (1979). Quantitative methods for analyzing travel behavior of individuals: Some recent developments. In D. A. Hensher and P. R. Stopher (eds.), *Behavioral Travel Modelling*. London: Croom Helm.
- McFadden, D. (1982). Qualitative response models. In W. Hildebrand (ed.), *Advances in Econometrics*. Cambridge University Press.
- McFadden, D. (1984). Econometric analysis of qualitative response models. In Z. Griliches and M. D. Intrilligator (eds.), *Handbook of Econometrics, Volume III*. Elsevier Science Publishers BV.
- Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, Vol. 78, No. 3: 691–692.
- Nerlove, M. and Press, S. J. (1973). Univariate and multivariate loglinear and logistic models. Rand Report No R-1306EDA/NIH.
- Peduzzi, P. N., Holford, T. R., and Hardy, R. J. (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics*, 36, 511–516.
- Pregibon, D. (1981). Logistic regression diagnostics. *Annals of Statistics*, 9, 705–724.
- Prentice, R. and Breslow, N. (1978). Retrospective studies and failure time models. *Biometrika*, 65, 153–158.
- Prentice, R. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66, 403–412.
- Rao, C. R. (1973). *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley & Sons.

- Santer, T. J. and Duffy, D. E. (2004). *The statistical analysis of discrete data*. New York: Springer-Verlag.
- \*Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6, 461-464.
- Steinberg, D. (1987). Interpretation and diagnostics of the multinomial and binary logistic regression using PROC MLOGIT. SAS Users Group International, Proceedings of the Twelfth Annual Conference, 1071-1073, Cary, N.C.: SAS Institute.
- Steinberg, D. (1991). The common structure of discrete choice and conditional logistic regression models. Unpublished paper. Department of Economics, San Diego State University.
- \*Steinberg, D. and Cardell, N. S. (1987). Logistic regression on pooled choice based samples and samples missing the dependent variable. Proceedings of the Social Statistics Section. Alexandria, Va.: American Statistical Association, 158-160.
- Train, K. (1986). *Qualitative choice analysis*. Cambridge, Mass.: MIT Press.
- Williams, D. A. (1986). Interval estimation of the median lethal dose. *Biometrics*, 42, 641-645.
- Wrigley, N. (2002). *Categorical data analysis for geographers and environmental scientists*. Caldwell, N.J.: Blackburn Press.

(\* indicates additional reference.)



# Loglinear Models

Laszlo Engelman

Loglinear models are useful for analyzing relationships among the factors of a multiway frequency table. The loglinear procedure computes maximum likelihood estimates of the parameters of a loglinear model by using the Newton-Raphson method. For each user-specified model, a test of fit of the model is provided, along with observed and expected cell frequencies, estimates of the loglinear parameters ( $\lambda$ s), standard errors of the estimates, the ratio of each  $\lambda$  to its standard error, and multiplicative effects ( $\text{EXP}(\lambda)$ ).

For each cell, you can request its contribution to the Pearson chi-square or the likelihood-ratio chi-square. Deviates, standardized deviates, Freeman-Tukey deviates, and likelihood-ratio deviates are available to characterize departures of the observed values from expected values.

When searching for the best model, you can request tests after removing each first-order effect or interaction term one at a time individually or hierarchically (when a lower-order effect is removed, so are the higher order interaction terms containing it). The models need not be hierarchical.

A model can explain the frequencies well in most cells, but poorly in a few. LOGLIN uses Freeman-Tukey deviates to identify the most divergent cell, fit a model without it, and continue in a stepwise manner identifying other outlier cells that depart from your model.

You can specify cells that contain structural zeros (cells that are empty naturally or by design, not by sampling), and fit a model to the subset of cells that remain. A test of fit for such a model is often called a test of quasi-independence.

Resampling procedures are available in this feature.

## Statistical Background

Researchers fit loglinear models to the cell frequencies of a multiway table in order to describe relationships among the categorical variables that form the table.

To introduce loglinear models, recall how to calculate expected values for the Pearson chi-square statistic. The expected value for a cell in a row  $i$  and column  $j$  is ( $F_{ij}$ ):

$$F_{ij} = \text{total count } (n) * (\text{proportion in row } i(p_i)) * (\text{proportion in column } j(p_j))$$

(Part of each expected value comes from the row it is in and part from the column it is in.) Now, by taking the log, we get an expression of the type:

$$\ln F_{ij} = \text{constant} + A_i + B_j$$

Thus the logarithm of the expected frequency is linear in certain parameters. Similarly, the loglinear model expresses the logarithm of the expected cell frequency as a linear function of these parameters in a manner analogous to that of analysis of variance.

In the above model, the expected value is computed under the null hypothesis of independence (that is, there is no interaction between the table factors). If this hypothesis is rejected, you would need more information than  $A_i$  and  $B_j$ . In fact, the usual chi-square test can be expressed as a test that the interaction term is needed in a model that estimates the log of the cell frequencies. We write this model as:

$$\ln F_{ij} = \text{constant} + A_i + B_j + AB_{ij}$$

or more commonly as:

$$\ln F_{ij} = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$$

where  $\theta$  is an overall mean effect and the parameters  $\lambda$  sum to zero over the levels of the row factors and the column factors. For a particular cell in a three-way table (a cell in the  $i$  row,  $j$  column, and  $k$  level of the third factor) we write:

$$\ln F_{ijk} = \theta + \lambda_i^A + \lambda_j^B + \lambda_k^C + \lambda_{ij}^{AB} + \lambda_{ik}^{AC} + \lambda_{jk}^{BC} + \lambda_{ijk}^{ABC}$$

The order of the effect is the number of indices in the subscript.

Notation in publications for loglinear model parameters varies. Grant Blank summarizes:

**SYSTAT                      FATHER + SON + FATHER \* SON**

Agresti (1984)               $\log m_{ij} = \mu + \lambda_i^F + \lambda_j^S + \lambda_{ij}^{FS}$

Fienberg (1980)               $\log m_{ij} = \mu + \mu_{1(i)} + \mu_{2(j)} + \mu_{12(ij)}$

Goodman (1978)               $\xi_{ij} = \theta + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$

Haberman (1978)               $\log m_{ij} = \mu + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB}$

Knoke and Burke (1980)       $G_{ij} = \theta + \lambda_i^F + \lambda_j^S + \lambda_{ij}^{FS}$

or, in multiplicative form,  $F_{ij} = \eta r_i^A r_j^B r_{ij}^{AB}$  where  $\xi_{ij} = \log(F_{ij})$ ,  $\theta = \log \eta$ ,  $\lambda_i^A = \log(r_i^A)$ , etc.  
Goodman (1971)

An important distinction between ANOVA and loglinear modeling is that in the latter, the focus is on the need for interaction terms; while in ANOVA, testing for main effects is the primary interest. Look back at the loglinear model for the two-way table—the usual chi-square tests the need for the  $AB_{ij}$  interaction, not for A alone or B alone.

The above loglinear model for a three-way table is saturated because it contains all possible terms or effects. Various smaller models can be formed by including only selected combinations of effects (or equivalently testing that certain effects are 0). An important goal in loglinear modeling is parsimony—that is, to see how few effects are needed to estimate the cell frequencies. You usually don't want to test that the main effect of a factor is 0 because this is the same as testing that the total frequencies are equal for all levels of the factor. For example, a test that the main effect for *SURVIVE\$* (alive, dead) is 0 simply tests whether the total number of survivors equals the number of nonsurvivors. If no interaction terms are included and the test is not significant (that is, the model fits), you can report that the table factors are independent. When there are more than two second-order effects, the test of an interaction is conditional on the other interactions and may not have a simple interpretation.

## ***Fitting a Loglinear Model***

To fit a loglinear model:

- First, screen for an appropriate model to test.



- Test the model, and if significant, compare its results with those for models with one or more additional terms. If not significant, compare results with models with fewer terms.
- For the model you select as best, examine fitted values and residuals, looking for cells (or layers within the table) with large differences between observed and expected (fitted) cell counts.

How do you determine which effects or terms to include in your loglinear model? Ideally, by using your knowledge of the subject matter of your study, you have a specific model in mind—that is, you want to make statements regarding the independence of certain table factors. Otherwise, you may want to screen for effects.

The likelihood-ratio chi-square is additive under partitioning for nested models. Two models are nested if all the effects of the first are a subset of the second. The likelihood ratio chi-square is additive because the statistic for the second model can be subtracted from that of the first. The difference provides a test of the additional effects—that is, the difference in the two statistics has an asymptotic chi-square distribution with degrees of freedom equal to the difference between those for the two model chi-squares (or the difference between the number of effects in the two models). This property does not hold for the Pearson chi-square. The additive property for the likelihood ratio chi-square is useful for screening effects to include in a model.

If you are doing exploratory research and lack firm knowledge about which effects to include, some statisticians suggest a strategy of starting with a large model and, step by step, identifying effects to delete. (You compare each smaller model nested within the larger one as described above.) But we *caution you about multiple testing*. If you test many models in a search for your ideal model, remember that the *p-value* associated with a specific test is valid when you execute one and only one test. That is, use *p-values* as relative measures when you test several models.

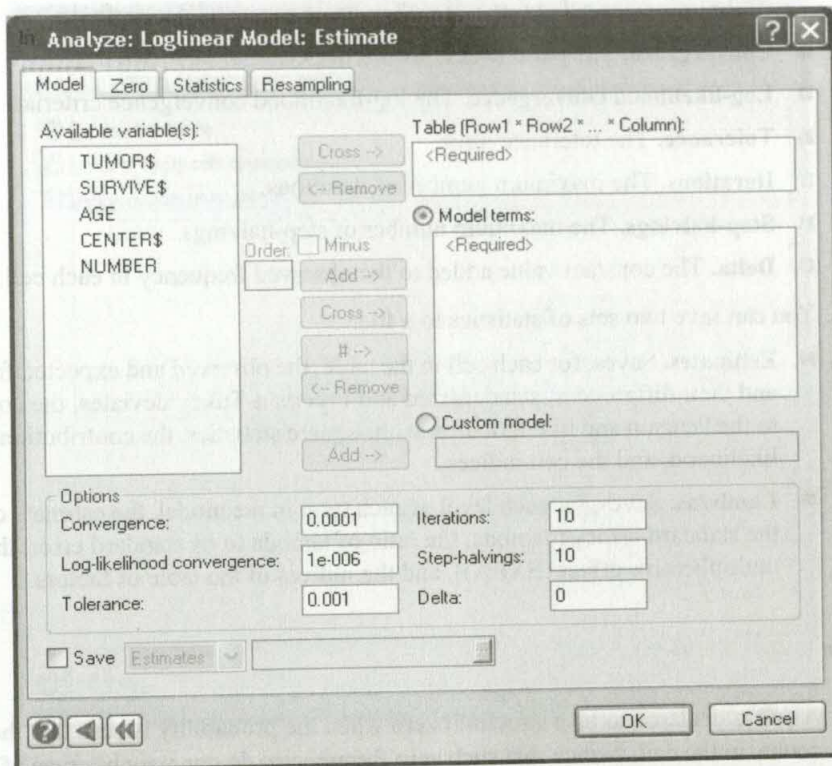
## ***Loglinear Models in SYSTAT***

### ***Loglinear Model: Estimate Dialog Box***

To open the Loglinear Model: Estimate dialog box, from the menus choose:

Analyze  
Loglinear Model  
Estimate...





The following must be specified:

**Model terms.** Build the model components (main effects and interactions) by adding terms to the Model terms text box. All variables should be categorical (either numerical or string). Click Cross to add interactions. Click # to include lower order effects with the interaction term, that is,  $A \# B = A + B + A * B$ . Check the Minus option with a selection of variables to remove (subset or all) model terms from previously defined model terms. The model terms can be defined up to a desired higher level of interaction using Order option. For example,  $(A + B + C)^2 = A + B + C + A * B + A * C + B * C$ .

**Custom Model.** Any valid loglinear model expression can be constructed using variable names and symbols: +, -, \*, #, ^, . For example,  $(A + B + C)^2 - (A \# B)$

**Table.** The variables that define the frequency table. Variables that are used in the model terms must be included in the frequency table.

The following optional computational controls can also be specified:

- **Convergence.** The parameter convergence criteria.
- **Log-likelihood convergence.** The log-likelihood convergence criteria.
- **Tolerance.** The tolerance limit.
- **Iterations.** The maximum number of iterations.
- **Step-halvings.** The maximum number of step-halvings.
- **Delta.** The constant value added to the observed frequency in each cell.

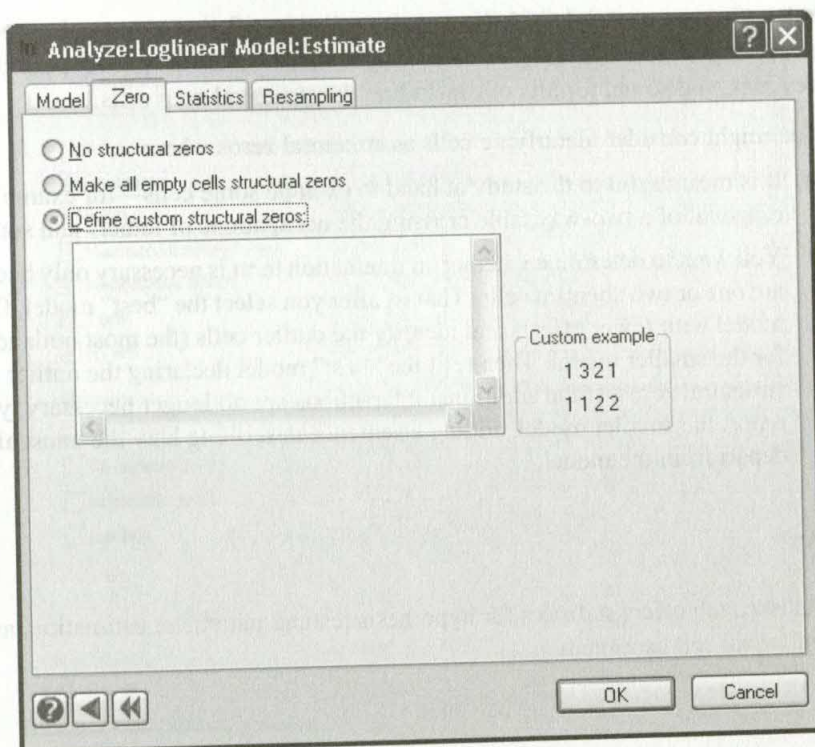
You can save two sets of statistics to a file:

- **Estimates.** Saves, for each cell in the table, the observed and expected frequencies and their differences, standardized and Freeman-Tukey deviates, the contribution to the Pearson and likelihood-ratio chi-square statistics, the contribution to the log-likelihood, and the cell indices.
- **Lambdas.** Saves, for each level of each term in the model, the estimate of lambda, the standard error of lambda, the ratio of lambda to its standard error, the multiplicative effect ( $\text{EXP}(\lambda)$ ), and the indices of the table of factors.

## Zero

A cell is declared to be a structural zero when the probability is zero that there are counts in the cell. Notice that such zero frequencies do not arise because of small samples but because the cells are empty naturally (a male hysterectomy patient) or by design (the diagonal of a two-way table comparing father's (rows) and son's (columns) occupations is not of interest when studying changes or mobility). A model can then be fit to the subset of cells that remain. A test of fit for such a model is often called a test of quasi-independence.

To specify structural zeros, click the Zero tab in the Analyze:Loglinear Model: Estimate dialog box.



The following can be specified:

**No structural zeros.** No cells are treated as structural zeros.

**Make all empty cells structural zeros.** Treats all empty cells with zero frequency as structural zeros. In the output, the corresponding cell which is defined as structural zero will be represented with an asterisk (\*) by default. If you give an option BLANK, these cells will be shown as blank cells. This option can be given only through commands.

**Define custom structural zeros.** Specifies one or more cells for treatment as structural zeros. List the index ( $n_1, n_2, \dots$ ) of each factor in the order in which the factor appears in the table. If you want to select a layer or level of a factor, use 0's for the other factors when specifying the indices. For example, in a table with four factors (*TUMOR\$* being the fourth factor), to declare the third level of *TUMOR\$* as structural zeros, use 0 0 0 3. Alternatively, you can replace the 0's with periods (... 3).



When fitting a model, LOGLIN excludes cells identified as structural zeros, and then, as in a regression analysis with zero weight cases, it can compute expected values, deviates, and so on, for all cells including the structural zero cells.

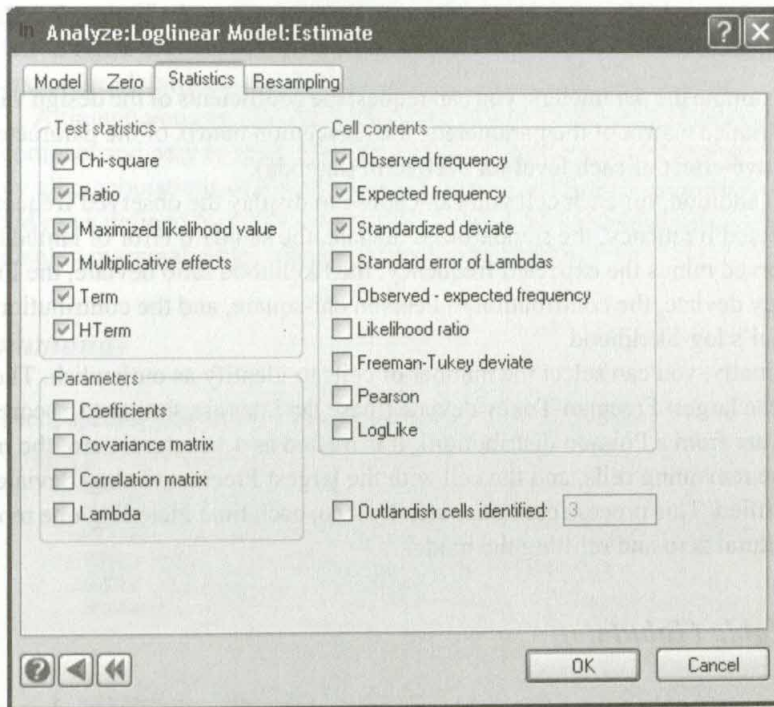
You might consider identifying cells as structural zeros when:

- It is meaningful to the study at hand to exclude some cells—for example, the diagonal of a two-way table crossing the occupations of fathers and sons.
- You want to determine whether an interaction term is necessary only because there are one or two aberrant cells. That is, after you select the “best” model, fit a second model with fewer effects and identify the outlier cells (the most outlandish cells) for the smaller model. Then refit the “best” model declaring the outlier cells to be structural zeros. If the additional interactions are no longer necessary, you might report the smaller model, adding a sentence describing how the unusual cell(s) depart from the model.

### *Statistics*

Statistics tab offers statistics for hypothesis testing, parameter estimation, and individual cell examination.





The following statistics are available:

- **Chi-square.** Displays Pearson and likelihood-ratio chi-square statistics for lack of fit.
- **Ratio.** Displays lambda divided by standard error of lambda. For large samples, this ratio can be interpreted as a standard normal deviate ( $z$  score).
- **Maximized likelihood value.** The log of the model's maximum likelihood value.
- **Multiplicative effects.** Multiplicative parameters,  $\text{EXP}(\lambda)$ . Large values indicate an increased probability for that combination of indices.
- **Term.** One at a time, LOGLIN removes each first-order effect and each interaction term from the model. For each smaller model, LOGLIN provides a likelihood-ratio chi-square for testing the fit of the model and the difference in the chi-square statistics between the smaller model and the full model.
- **HTerm.** Tests each term by removing it and its higher order interactions from the model. These tests are similar to those in Term except that only hierarchical models

are tested—if a lower-order effect is removed, so are the higher-order effects that include it.

To examine the parameters, you can request the coefficients of the design variables, the covariance matrix of the parameters, the correlation matrix of the parameters, and the additive effect of each level for each term (lambda).

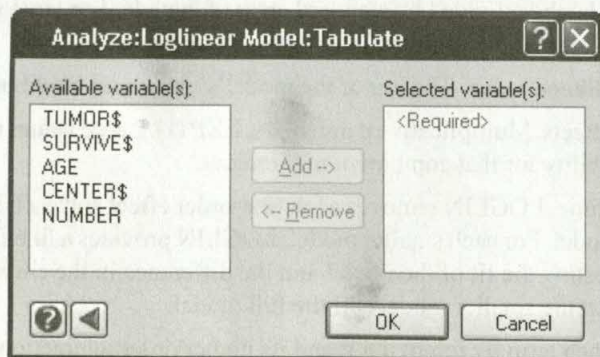
In addition, for each cell you can choose to display the observed frequency, the expected frequency, the standardized deviate, the standard error of lambda, the observed minus the expected frequency, the likelihood ratio deviate, the Freeman-Tukey deviate, the contribution to Pearson chi-square, and the contribution to the model's log-likelihood.

Finally, you can select the number of cells to identify as outlandish. The first cell has the largest Freeman-Tukey deviate (these deviates are similar to  $z$  scores when the data are from a Poisson distribution). It is treated as a structural zero, the model is fit to the remaining cells, and the cell with the largest Freeman-Tukey deviate is identified. This process continues step by step, each time including one more cell as a structural zero and refitting the model.

## Frequency Table (Tabulate)

If you want only a frequency table and no analysis, from the menus choose:

Analyze  
Loglinear Model  
Tabulate...



Simply specify the table factors in the same order in which you want to view them from left to right. In other words, the last variable selected defines the columns of the table and cross-classifications of all preceding variables define the rows.

Although you can also form multiway tables, tables for loglinear models are more compact and easy to read. Multiway tables form a series of two-way tables stratified by all combinations of the other table factors. Loglinear models create one table, with the rows defined by factor combinations. However, loglinear model tables do not display marginal totals, whereas Multiway tables do.

## Using Commands

First, specify your data with `USE filename`. Continue with:

```

LOGLIN
FREQ var
TABULATE var1*var2*...
MODEL variables defining table = terms of model
ZERO CELL n1, n2, ...or Empty/ BLANK
SAVE filename / ESTIMATES or LAMBDA
PLENGTH SHORT or MEDIUM or LONG or NONE,
      / OBSFREQ CHISQ RATIO MLE EXPECT STAND ELAMBDA,
      TERM HTERM PARAM COVA CORR LAMBDA SELAMBDA DEVIATES,
      LRDEV FTDEV PEARSON LOGLIKE CELLS=n
ESTIMATE / DELTA=n LCONV=n CONV=n TOL=n ITER=n HALF=n
SAMPLE = BOOT(m,n)
        = JACK
        = SIMPLE(m,n)

```

## Usage Considerations

**Types of data.** LOGLIN uses a cases-by-variables rectangular file or data recorded as frequencies with cell indices.

**Print options.** You can control what report panels appear in the output by globally setting output length to SHORT, MEDIUM, or LONG. You can also use the PLENGTH command in LOGLIN to request reports individually. You can specify individual panels by specifying the particular option.

Short output panels include the observed frequency for each cell, the Pearson and likelihood-ratio chi-square statistics, lambdas divided by their standard errors, the log of the model's maximized likelihood value, and a report of the three most outlandish cells.

Medium results include all of the above, plus the following: the expected frequency for each cell (current model), standardized deviations, multiplicative effects, a test of



each term by removing it from the model, a test of each term by removing it and its higher-order interactions from the model, and the five most outlandish cells.

Long results add the following: coefficients of design variables, the covariance matrix of the parameters, the correlation matrix of the parameters, the additive effect of each level for each term, the standard errors of the lambdas, the observed minus the expected frequency for each cell, the contribution to the Pearson chi-square from each cell, the likelihood-ratio deviate for each cell, the Freeman-Tukey deviate for each cell, the contribution to the model's log-likelihood from each cell, and the 10 most outlandish cells.

As a PLENGTH option, you can also specify CELLS=*n*, where *n* is the number of outlandish cells to identify.

**Quick Graphs.** LOGLIN produces no Quick Graphs.

**Saving files.** For each level of a term included in your model, you can save the estimate of lambda, the standard error of lambda, the ratio of lambda to its standard error, the multiplicative effect, and the marginal indices of the effect. Alternatively, for each cell, you can save the observed and expected frequencies, its deviates (listed above), the Pearson and likelihood-ratio chi-square, the contributions to the log-likelihood, and the cell indices.

**BY groups.** LOGLIN analyzes each level of any BY variables separately.

**Case frequencies.** LOGLIN uses the FREQ variable, if present, to duplicate cases.

**Case weights.** WEIGHT variables have no effect in LOGLIN.



## Examples

### Example 1

#### Loglinear Modeling of a Four-Way Table

In this example, we use the Morrison breast cancer data stored in the *CANCER* data file (Bishop, Fienberg and Holland, 1977) and treat the data as a four-way frequency table:

*CENTERS* Center or city where the data were collected  
*SURVIVE\$* Survival—dead or alive  
*AGE* Age groups of under 50, 50 to 69, and 70 or over  
*TUMOR\$* Tumor diagnosis (called INFLAPP by some researchers) with levels:  
 –Minimal inflammation and benign  
 –Greater inflammation and benign  
 –Minimal inflammation and malignant  
 –Greater inflammation and malignant

The *CANCER* data include one record for each of the 72 cells formed by the four table factors. Each record includes a variable, *NUMBER*, that has the number of women in the cell plus numeric or character value codes to identify the levels of the four factors that define the cell.

For the first model of the *CANCER* data, you include three two-way interactions.

The input is:

```
USE CANCER
LOGLIN
  FREQ number
  LABEL age / 50='Under 50', 60='50 to 69', 70='70 & Over'
  ORDER center$ survive$ tumor$ / SORT=NONE
  MODEL center$*age*survive$*tumor$ = center$ + age,
                                         + survive$ + tumor$,
                                         + age*center$,
                                         + survive$*center$,
                                         + tumor$*center$

  PLENGTH SHORT / EXPECT LAMBDA
  ESTIMATE / DELTA=0.5
```

The MODEL statement has two parts: table factors and terms (effects to fit). Table factors appear to the left of the equal sign and terms are on the right. The layout of the table is determined by the order in which the variables are specified—for example, specify *TUMOR\$* last so its levels determine the columns.

The LABEL statement assigns category names to the numeric codes for AGE. If the statement is omitted, the data values label the categories. By default, SYSTAT orders string variables alphabetically, so we specify SORT = NONE to list the categories for the other factors as they first appear in the data file.

We specify DELTA = 0.5 to add 0.5 to each cell frequency. This option is common in multiway table procedures as an aid when some cell sizes are sparse. It is of little use in practice and is used here only to make the results compare with those reported elsewhere.

The output is:

Case frequencies determined by value of variable NUMBER

Number of Cells (product of levels) : 72  
Total count : 764

#### Observed Frequencies

CENTERS\$	AGE	SURVIVES\$	TUMOR\$			
			MinMalig	MinBengn	MaxMalig	MaxBengn
Tokyo	Under 50	Dead	9.000	7.000	4.000	3.000
		Alive	26.000	68.000	25.000	9.000
	50 to 69	Dead	9.000	9.000	11.000	2.000
		Alive	20.000	46.000	18.000	5.000
	70 & Over	Dead	2.000	3.000	1.000	0.000
		Alive	1.000	6.000	5.000	1.000
Boston	Under 50	Dead	6.000	7.000	6.000	0.000
		Alive	11.000	24.000	4.000	0.000
	50 to 69	Dead	8.000	20.000	3.000	2.000
		Alive	18.000	58.000	10.000	3.000
	70 & Over	Dead	9.000	18.000	3.000	0.000
		Alive	15.000	26.000	1.000	1.000
Glamorgn	Under 50	Dead	16.000	7.000	3.000	0.000
		Alive	16.000	20.000	8.000	1.000
	50 to 69	Dead	14.000	12.000	3.000	0.000
		Alive	27.000	39.000	10.000	4.000
	70 & Over	Dead	3.000	7.000	3.000	0.000
		Alive	12.000	11.000	4.000	1.000

Pearson Chi-square : 57.527 df : 51 p-value : 0.246  
LR Chi-square : 55.833 df : 51 p-value : 0.298  
Raftery's BIC : -282.734  
Dissimilarity : 9.953

## Expected Values

CENTERS	AGE	SURVIVE\$	TUMORS			
			MinMalig	MinBengn	MaxMalig	MaxBengn
Tokyo	Under 50	Dead	7.852	15.928	7.515	2.580
		Alive	28.076	56.953	26.872	9.225
	50 to 69	Dead	6.281	12.742	6.012	2.064
		Alive	22.460	45.563	21.498	7.380
	70 & Over	Dead	1.165	2.363	1.115	0.383
		Alive	4.166	8.451	3.988	1.369
Boston	Under 50	Dead	5.439	12.120	2.331	0.699
		Alive	10.939	24.378	4.688	1.406
	50 to 69	Dead	11.052	24.631	4.737	1.421
		Alive	22.231	49.542	9.527	2.858
	70 & Over	Dead	6.754	15.052	2.895	0.868
		Alive	13.585	30.276	5.822	1.747
Glamorgn	Under 50	Dead	9.303	10.121	3.476	0.920
		Alive	19.989	21.746	7.468	1.977
	50 to 69	Dead	14.017	15.249	5.237	1.386
		Alive	30.117	32.764	11.252	2.979
	70 & Over	Dead	5.582	6.073	2.086	0.552
		Alive	11.993	13.048	4.481	1.186

## Log-Linear Effects (Lambda)

THETA

1.826

CENTERS		
Tokyo	Boston	Glamorgn

0.049	0.001	-0.050
-------	-------	--------

AGE		
Under 50	50 to 69	70 & Over
0.145	0.444	-0.589

SURVIVES	
Dead	Alive
-0.456	0.456

TUMORS			
MinMalig	MinBengn	MaxMalig	MaxBengn
0.480	1.011	-0.145	-1.346

CENTERS	AGE		
	Under 50	50 to 69	70 & Over
Tokyo	0.565	0.043	-0.609
Boston	-0.454	-0.043	0.497
Glamorgn	-0.111	0.000	0.112

CENTERS	SURVIVES	
	Dead	Alive
Tokyo	-0.181	0.181
Boston	0.107	-0.107
Glamorgn	0.074	-0.074

CENTERS\$	TUMORS\$			
	MinMalig	MinBegn	MaxMalig	MaxBegn
Tokyo	-0.368	-0.191	0.214	0.345
Boston	0.044	0.315	-0.178	-0.181
Glamorgn	0.323	-0.123	-0.036	-0.164

## Standardized Parameter Estimates (Lambda / Standard Error of Lambda)

## THETA

30.528

Tokyo	CENTERS\$	
	Boston	Glamorgn
0.596	0.014	-0.586

Under 50	AGE	
	50 to 69	70 & Over
2.627	8.633	-8.649

Dead	SURVIVES\$	
	Dead	Alive
-11.548	11.548	

MinMalig	TUMORS\$		
	MinBegn	MaxMalig	MaxBegn
6.775	15.730	-1.718	-10.150

CENTERS\$	AGE		
	Under 50	50 to 69	70 & Over
Tokyo	7.348	0.576	-5.648
Boston	-5.755	-0.618	5.757
Glamorgn	-1.418	-0.003	1.194

CENTERS\$	SURVIVES\$	
	Dead	Alive
Tokyo	-3.207	3.207
Boston	1.959	-1.959
Glamorgn	1.304	-1.304

CENTERS\$	TUMORS\$			
	MinMalig	MinBegn	MaxMalig	MaxBegn
Tokyo	-3.862	-2.292	2.012	2.121
Boston	0.425	3.385	-1.400	-0.910
Glamorgn	3.199	-1.287	-0.289	-0.827

Model ln(MLE) | -160.563

## The 3 most Outlandish Cells (based on FTD, stepwise)

ln(MLE)	LR Chi-square	p-value	Frequency	CENTERS\$	AGE	SURVIVES\$	TUMORS\$
-154.685	11.755	0.001	7	1	1	1	2
-150.685	8.001	0.005	1	2	3	2	3
-145.024	11.321	0.001	16	3	1	1	1

Initially, SYSTAT produces a frequency table for the data. We entered cases for 72 cells. The total frequency count across these cells is 764—that is, there are 764 women in the sample. Notice that the order of the factors is the same order we specified in the



MODEL statement. The last variable (*TUMORS*) defines the columns; the remaining variables define the rows.

The test of fit is not significant for either the Pearson chi-square or the likelihood-ratio test, indicating that your model with its three two-way interactions does not disagree with the observed frequencies. The model statement describes an association between study center and age, survival, and tumor status. However, at each center, the other three factors are independent. Because the overall goal is parsimony, we could explore whether any of the interactions can be dropped.

Raftery's BIC (Bayesian Information Criterion) adjusts the chi-square for both the complexity of the model (measured by degrees of freedom) and the size of the sample. It is the likelihood-ratio chi-square minus the degrees of freedom for the current model times the natural log of the sample size. If BIC is negative, you can conclude that the model is preferable to the saturated model. When comparing alternative models, select the model with the lowest BIC value.

The index of dissimilarity can be interpreted as the percentage of cases that need to be relocated in order to make the observed and expected counts equal. For these data, you would have to move about 9.95% of the cases to make the expected frequencies fit.

The expected frequencies are obtained by fitting the loglinear model to the observed frequencies. Compare these values with the observed frequencies. Values for corresponding cells will be similar if the model fits well.

After the expected values, SYSTAT lists the parameter estimates for the model you requested. Usually, it is of more interest to examine these estimates divided by their standard errors. Here, however, we display them in order to relate them to the expected values. For example, the observed frequency for the cell in the upper left corner (*Tokyo, Under 50, Dead, MinMalig*) is 9. To find the expected frequency under your model, you add the estimates (from each panel, select the term that corresponds to your cell):

theta	1.826	C*A	0.565
CENTER\$	0.049	C*S	-0.181
AGE	0.145	C*T	-0.368
SURVIVES	-0.456		
TUMORS	0.480		

and then use SYSTAT's calculator to sum the estimates:

$$\text{CALC } 1.826 + 0.049 + 0.145 - 0.456 + 0.480 + 0.565 - 0.181 - 0.368$$

and SYSTAT responds 2.06. Take the antilog of this value:

$$\text{CALC EXP}(2.06)$$

and SYSTAT responds 7.846. In the panel of expected values, this number is printed as 7.852 (in its calculations, SYSTAT uses more digits following the decimal point). Thus, for this cell, the sample includes 9 women (observed frequency) and the model predicts 7.85 women (expected frequency).

The ratio of the parameter estimates to their asymptotic standard errors is part of the default output. Examine these values to better understand the relationships among the table factors. Because, for large samples, this ratio can be interpreted as a standard normal deviate ( $z$  score), you can use it to indicate significant parameters—for example, for an interaction term, significant positive (or negative) associations. In the *CENTER\$* by *AGE* panel, the ratio for young women from Tokyo is very large (7.348), implying a significant positive association, and that for older Tokyo women is extremely negative (−5.648). The reverse is true for the women from Boston. If you use the Column Percent option in XTAB to print column percentages for *CENTER\$* by *AGE*, you will see that among the women under 50, more than 50% are from Tokyo (53.9), while only 20.7% are from Boston. In the 70 and over age group, 14% are from Tokyo and 55% are from Boston.

The *Alive* estimate for Tokyo shows a strong positive association (3.207) with survival in Tokyo. The relationship in Boston is negative (−1.959). In this study, the overall survival rate is 72.5%. In Tokyo, 79.3% of the women survived, while in Boston, 67.6% survived. There is a negative association for having a malignant tumor with minimal inflammation in Tokyo (−3.862). The same relationship is strongly positive in Glamorgan (3.199).

Cells that depart from the current model are identified as outlandish in a stepwise manner. The first cell has the largest Freeman-Tukey deviate (these deviates are similar to  $z$  scores when the data are from a Poisson distribution). It is treated as a structural zero, the model is fit to the remaining cells, and the cell with the largest Freeman-Tukey deviate is identified. This process continues step by step, each time including one more cell as a structural zero and refitting the model.

For the current model, the observations in the cell corresponding to the youngest nonsurvivors from Tokyo with benign tumors and minimal inflammation (*Tokyo, Under 50, Dead, MinBengn*) differs the most from its expected value. There are 7 women in the cell and the expected value is 15.9 women. The next most unusual cell is 2,3,2,3 (*Boston, 70 & Over, Alive, MaxMalig*), and so on.

*Medium Output*

We continue the previous analysis, repeating the same model, but changing the PLENGTH (output length) setting to request medium-length results:

The input is:

```
USE CANCER
LOGLIN
  FREQ number
  LABEL age / 50='Under 50', 60='50 to 69', 70='70 & Over'
  ORDER center$ survive$ tumor$ / SORT=NONE
  MODEL center$*age*survive$*tumor$ = age # center$,
    + survive$ # center$,
    + tumor$ # center$

  PLENGTH MEDIUM
  ESTIMATE / DELTA = 0.5
```

Notice that we use shortcut notation to specify the model.

The output is:

Standardized Deviates = (Obs-Exp)/sqrt(Exp)

CENTER\$	AGE	SURVIVE\$	TUMOR\$			
			MinMalig	MinBengn	MaxMalig	MaxBengn
Tokyo	Under 50	Dead	0.410	-2.237	-1.282	0.262
		Alive	-0.392	1.464	-0.361	-0.074
	50 to 69	Dead	1.085	-1.048	2.034	-0.044
		Alive	-0.519	0.065	-0.754	-0.876
	70 & Over	Dead	0.774	0.414	-0.109	-0.619
		Alive	-1.551	-0.843	0.507	-0.315
Boston	Under 50	Dead	0.241	-1.471	2.403	-0.836
		Alive	0.018	-0.077	-0.318	-1.186
	50 to 69	Dead	-0.918	-0.933	-0.798	0.486
		Alive	-0.897	1.202	0.153	0.084
	70 & Over	Dead	0.864	0.760	0.062	-0.932
		Alive	0.384	-0.777	-1.999	-0.565
Glamorgn	Under 50	Dead	2.196	-0.981	-0.255	-0.959
		Alive	-0.892	-0.374	0.195	-0.695
	50 to 69	Dead	-0.004	-0.832	-0.977	-1.177
		Alive	-0.568	1.089	-0.373	0.592
	70 & Over	Dead	-1.093	0.376	0.633	-0.743
		Alive	0.002	-0.567	-0.227	-0.171

Multiplicative Effects = exp(Lambda)

THETA			
6.209			
AGE			
Under 50	50 to 69	70 & Over	
1.156	1.559	0.555	



CENTERS\$		
Tokyo	Boston	Glamorgn
1.050	1.001	0.951

SURVIVES\$	
Dead	Alive
0.634	1.578

TUMOR\$			
MinMalig	MinBengn	MaxMalig	MaxBengn
1.616	2.748	0.865	0.260

CENTERS\$	AGE			
	Under 50	50 to 69	70 & Over	
Tokyo	1.760	1.044	0.544	
Boston	0.635	0.958	1.644	
Glamorgn	0.895	1.000	1.118	

CENTER\$	SURVIVES\$	
	Dead	Alive
Tokyo	0.835	1.198
Boston	1.113	0.899
Glamorgn	1.077	0.929

CENTERS\$	TUMOR\$			
	MinMalig	MinBengn	MaxMalig	MaxBengn
Tokyo	0.692	0.826	1.238	1.412
Boston	1.045	1.370	0.837	0.834
Glamorgn	1.382	0.884	0.965	0.849

Model ln(MLE) : -160.563

#### Tests for Model Terms

Term Tested	The Model without the Term				Removal of Term from Model			
	ln(MLE)	Chi-square	df	p-value	Chi-square	df	p-value	
AGE	-216.120	166.946	53	0.000	111.114	2	0.000	
CENTERS\$	-160.799	56.306	53	0.352	0.473	2	0.789	
SURVIVES\$	-234.265	203.238	52	0.000	147.405	1	0.000	
TUMOR\$	-344.471	423.649	54	0.000	367.817	3	0.000	
CENTERS\$*AGE	-196.672	128.050	55	0.000	72.217	4	0.000	
CENTERS\$*SURVIVES\$	-166.007	66.721	53	0.097	10.888	2	0.004	
CENTERS\$*TUMOR\$	-178.267	91.241	57	0.003	35.408	6	0.000	

#### Tests for Hierarchical Terms

Term Tested Hierarchically	The Model without the Term				Removal of Term from Model			
	ln(MLE)	Chi-square	df	p-value	Chi-square	df	p-value	
AGE	-246.779	228.264	57	0.000	172.432	6	0.000	
CENTERS\$	-224.289	183.285	65	0.000	127.453	14	0.000	
SURVIVES\$	-242.434	219.574	54	0.000	163.741	3	0.000	
TUMOR\$	-363.341	461.390	60	0.000	405.557	9	0.000	

#### The 5 most Outlandish Cells (based on FTD, stepwise)

ln(MLE)	LR Chi-square	p-value	Frequency	CENTERS\$	AGE	SURVIVES\$	TUMOR\$
-154.685	11.755	0.001	7	1	1	1	2
-150.685	8.001	0.005	1	2	3	2	3
-145.024	11.321	0.001	16	3	1	1	1
-140.740	8.569	0.003	6	2	1	1	3
-136.662	8.157	0.004	11	1	2	1	3



The goodness-of-fit tests provide an *overall* indication of how close the expected values are to the cell counts. Just as you study residuals for each case in multiple regression, you can use deviates to compare the observed and expected values for each cell. A standardized deviate is the square root of each cell's contribution to the Pearson chi-square statistic—that is, (the observed frequency minus the expected frequency) divided by the square root of the expected frequency. These values are similar to  $z$  scores. For the second cell in the first row, the expected value under your model is considerably larger than the observed count (its deviate is  $-2.237$ , the observed count is 7, and the expected count is 15.9). Previously, this cell was identified as the most outlandish cell using Freeman-Tukey deviates.

Note that LOGLIN produces five types of deviates or residuals: standardized, the observed minus the expected frequency, the likelihood-ratio deviate, the Freeman-Tukey deviate, and the Pearson deviate.

Estimates of the multiplicative parameters equal  $\text{Exp}(\lambda)$ . Look for values that depart markedly from 1.0. Very large values indicate an increased probability for that combination of indices and, conversely, a value considerably less than 1.0 indicates an unlikely combination. A test of the hypothesis that a multiplicative parameter equals 1.0 is the same as that for  $\lambda$  equal to 0; so use the values of  $(\lambda)/SE$  to test the values in this panel. For the *CENTER\$* by *AGE* interaction, the most likely combination is women under 50 from Tokyo (1.76); the least likely combination is women 70 and over from Tokyo (0.544).

After listing the multiplicative effects, SYSTAT tests reduced models by removing each first-order effect and each interaction from the model one at a time. For each smaller model, LOGLIN provides:

- A likelihood-ratio chi-square for testing the fit of the model
- The difference in the chi-square statistics between the smaller model and the full model

The likelihood-ratio chi-square for the full model is 55.833. For a model that omits *AGE*, the likelihood-ratio chi-square is 166.95. This smaller model does not fit the observed frequencies ( $p\text{-value} < 0.00005$ ). To determine whether the removal of this term results in a significant decrease in the fit, look at the difference in the statistics:  $166.95 - 55.833 = 111.117$ ,  $p\text{-value} < 0.00005$ . The fit worsens significantly when *AGE* is removed from the model.

From the second line in this panel, it appears that a model without the first-order term for *CENTER\$* fits ( $p\text{-value} = 0.3523$ ). However, removing any of the two-way interactions involving *CENTER\$* significantly decreases the model fit.

The hierarchical tests are similar to the preceding tests except that only hierarchical models are tested—if a lower-order effect is removed, so are the higher-order effects that include it. For example, in the first line, when *CENTER\$* is removed, the three interactions with *CENTER\$* are also removed. The reduction in the fit is significant ( $p\text{-value} < 0.00005$ ). Although removing the first-order effect of *CENTER\$* does not significantly alter the fit, removing the higher-order effects involving *CENTER\$* decreases the fit substantially.

## Example 2

### Screening Effects

In this example, you pretend that no models have been fit to the *CANCER* data (that is, you have not seen the other example). As a place to start, first fit a model with all second-order interactions finding that it fits. Then fit models nested within the first by using results from the *HTERM* (terms tested hierarchically) panel to guide your selection of terms to be removed.

Here's a summary of your instructions: you study the output generated from the first *MODEL* and *ESTIMATE* statements and decide to remove *AGE* by *TUMOR\$*. After seeing the results for this smaller model, you decide to remove *AGE* by *SURVIVE\$*, too. To carry out these steps, the input is:

```
USE CANCER
LOGLIN
  FREQ number
  PLENGTH NONE / CHI HTERM
  MODEL center$*age*survive$*tumor$ = tumor$..center$^2
  ESTIMATE / DELTA=0.5
  MODEL center$*age*survive$*tumor$ = tumor$..center$^2,
    - age*tumor$
  ESTIMATE / DELTA=0.5
  MODEL center$*age*survive$*tumor$ = tumor$..center$^2,
    - age*tumor$,
    - age*survive$
  ESTIMATE / DELTA=0.5
  MODEL center$*age*survive$*tumor$ = tumor$..center$^2,
    - age*tumor$,
    - age*survive$,
    - tumor$*survive$
  ESTIMATE / DELTA=0.5
```

The output is:

#### All two-way interactions

Pearson Chi-square : 40.165 df : 40 p-value : 0.463  
 LR Chi-square : 39.921 df : 40 p-value : 0.474  
 Raftery's BIC : -225.622  
 Dissimilarity : 7.643

#### Tests for Hierarchical Terms

Term Tested Hierarchically	The Model without ln(MLE)	the Term Chi-square	df	p-value	Removal of Term from Model Chi-square	df	p-value
TUMORS	-361.233	457.172	58	0.000	417.251	18	0.000
SURVIVES	-241.675	218.056	48	0.000	178.135	8	0.000
AGE	-241.668	218.043	54	0.000	178.122	14	0.000
CENTERS	-213.996	162.699	54	0.000	122.778	14	0.000
SURVIVES*TUMORS	-157.695	50.097	43	0.212	10.176	3	0.017
AGE*TUMORS	-153.343	41.393	46	0.665	1.473	6	0.961
AGE*SURVIVES	-154.693	44.093	42	0.383	4.173	2	0.124
CENTERS*TUMORS	-169.724	74.154	46	0.005	34.233	6	0.000
CENTERS*SURVIVES	-156.501	47.709	42	0.252	7.788	2	0.020
CENTERS*AGE	-186.011	106.728	44	0.000	66.808	4	0.000

#### Remove AGE \* TUMORS

Pearson Chi-square : 41.828 df : 46 p-value : 0.648  
 LR Chi-square : 41.393 df : 46 p-value : 0.665  
 Raftery's BIC : -263.981  
 Dissimilarity : 7.868

#### Tests for Hierarchical Terms

Term Tested Hierarchically	The Model without ln(MLE)	the Term Chi-square	df	p-value	Removal of Term from Model Chi-square	df	p-value
TUMORS	-361.233	457.172	58	0.000	415.779	12	0.000
SURVIVES	-242.434	219.574	54	0.000	178.181	8	0.000
AGE	-241.668	218.043	54	0.000	176.649	8	0.000
CENTERS	-215.687	166.082	60	0.000	124.688	14	0.000
SURVIVES*TUMORS	-158.454	51.615	49	0.372	10.221	3	0.017
AGE*SURVIVES	-155.452	45.611	48	0.571	4.218	2	0.121
CENTERS*TUMORS	-171.415	77.537	52	0.012	36.143	6	0.000
CENTERS*SURVIVES	-157.291	49.290	48	0.421	7.896	2	0.019
CENTERS*AGE	-187.702	110.111	50	0.000	68.718	4	0.000

#### Remove AGE \* TUMORS and AGE \* SURVIVES

Pearson Chi-square : 45.358 df : 48 p-value : 0.582  
 LR Chi-square : 45.611 df : 48 p-value : 0.571  
 Raftery's BIC : -273.040  
 Dissimilarity : 8.472

#### Tests for Hierarchical Terms

Term Tested Hierarchically	The Model without ln(MLE)	the Term Chi-square	df	p-value	Removal of Term from Model Chi-square	df	p-value
TUMORS	-363.341	461.390	60	0.000	415.779	12	0.000
SURVIVES	-242.434	219.574	54	0.000	173.963	6	0.000
AGE	-241.668	218.043	54	0.000	172.432	6	0.000
CENTERS	-219.546	173.799	62	0.000	128.188	14	0.000
SURVIVES*TUMORS	-160.563	55.833	51	0.298	10.221	3	0.017
CENTERS*TUMORS	-173.524	81.754	54	0.009	36.143	6	0.000
CENTERS*SURVIVES	-161.264	57.234	50	0.224	11.623	2	0.003
CENTERS*AGE	-191.561	117.828	52	0.000	72.217	4	0.000



Remove AGE \* TUMOR\$ , AGE \* SURVIVE\$ and TUMOR\$ \* SURVIVE\$

Pearson Chi-square : 57.527 df : 51 p-value : 0.246  
 LR Chi-square : 55.833 df : 51 p-value : 0.298  
 Raftery's BIC : -282.734  
 Dissimilarity : 9.953

#### Tests for Hierarchical Terms

Term Tested Hierarchically	ln (MLE)	The Model without Chi-square	the Term df	p-value	Removal of Term from Model Chi-square	df	p-value
TUMOR\$	-363.341	461.390	60	0.000	405.557	9	0.000
SURVIVE\$	-242.434	219.574	54	0.000	163.741	3	0.000
AGE	-246.779	228.264	57	0.000	172.432	6	0.000
CENTER\$	-224.289	183.285	65	0.000	127.453	14	0.000
CENTER\$*TUMOR\$	-178.267	91.241	57	0.003	35.408	6	0.000
CENTER\$*SURVIVE\$	-166.007	66.721	53	0.097	10.888	2	0.004
CENTER\$*AGE	-196.672	128.050	55	0.000	72.217	4	0.000

The likelihood-ratio chi-square for the model that includes all two-way interactions is 39.9 ( $p\text{-value} = 0.4738$ ). If the AGE by TUMOR\$ interaction is removed, the chi-square for the smaller model is 41.39 ( $p\text{-value} = 0.6654$ ). Does the removal of this interaction cause a significant change? No, chi-square = 1.47 ( $p\text{-value} = 0.9613$ ). This chi-square is computed as 41.39 minus 39.92 with 46 minus 40 degrees of freedom. The removal of this interaction results in the least change, so you remove it first. Notice also that the estimate of the maximized likelihood function is largest when this second-order effect is removed (-153.343).

The model chi-square for the second model is the same as that given for the first model with AGE \* TUMOR\$ removed (41.3934). Here, if AGE by SURVIVE\$ is removed, the new model fits ( $p\text{-value} = 0.5713$ ) and the change between the model minus one interaction and that minus two interactions is insignificant ( $p\text{-value} = 0.1214$ ).

If SURVIVE\$ by TUMOR\$ is removed from the current model with four interactions, the new model fits ( $p\text{-value} = 0.2981$ ). The change in fit is not significant ( $p\text{-value} = 0.0168$ ). Should we remove any other terms? Looking at the HTERM panel for the model with three interactions, you see that a model without CENTER\$ by SURVIVE\$ has a marginal fit ( $p\text{-value} = 0.0975$ ) and the chi-square for the difference is significant ( $p\text{-value} = 0.0043$ ). Although the goal is parsimony and technically a model with only two interactions does fit, you opt for the model that also includes CENTER\$ by SURVIVE\$ because it is a significant improvement over the very smallest model.



### Example 3 Structural Zeros

This example identifies outliers and then declares them to be structural zeros. You wonder if any of the interactions in the model that fit in the example on loglinear modeling for a four-way table are necessary only because of a few unusual cells. To identify the unusual cells, first pull back from your "ideal" model and fit a model with main effects only, asking for the four most unusual cells. (Why four cells? Because 5% of 72 cells is 3.6 or roughly 4).

The input is

```
USE CANCER
LOGLIN
  FREQ number
  ORDER center$ survive$ tumor$ / SORT=NONE
  MODEL center$*age*survive$*tumor$ = tumor$ .. center$
  PLENGTH SHORT / CELLS=4
  ESTIMATE / DELTA=0.5
```

Of course this model doesn't fit, but the following are selections from the output:

The output is:

```
Pearson Chi-square : 181.389   df : 63   p-value : 0.000
LR Chi-square      : 174.346   df : 63   p-value : 0.000
Raftery's BIC      : -243.884
Dissimilarity      : 19.385
```

#### The 4 most Outlandish Cells (based on FTD, stepwise)

ln(MLE)	LR Chi-square	p-value	Frequency	CENTER\$	AGE	SURVIVES\$	TUMOR\$
-203.261	33.118	0.000	68	1	1	2	2
-195.262	15.997	0.000	1	1	3	2	1
-183.471	23.582	0.000	25	1	1	2	3
-176.345	14.253	0.000	6	1	3	2	2

Next, fit your "ideal" model, identifying these four cells as structural zeros and also requesting PLENGTH SHORT / HTERM to test the need for each interaction term.

### Defining Four Cells As Structural Zeros

Continuing from the analysis of main effects only, now specify your original model with its three second-order effects.

The input for this is:

```
MODEL center$*age*survive$*tumor$ = ,
      (age + survive$ + tumor$) # center$
ZERO CELL=1 1 2 2 CELL=1 3 2 1 CELL=1 1 2 3 CELL=1 3 2 2
PLENGTH SHORT / HTERM
ESTIMATE / DELTA=0.5
```

The following are selections from the output. Notice that asterisks mark the structural zero cells.

The output is:

```
Number of Cells (product of levels) : 72
Number of structural zero cells      : 4
Total count                          : 664
```

#### Observed Frequencies

CENTERS\$	AGE	SURVIVES\$	TUMORS\$			
			MinMalig	MinBegn	MaxMalig	MaxBegn
Tokyo	50	Dead	9.000	7.000	4.000	3.000
		Alive	26.000	*68.000	*25.000	9.000
	60	Dead	9.000	9.000	11.000	2.000
		Alive	20.000	46.000	18.000	5.000
	70	Dead	2.000	3.000	1.000	0.000
		Alive	*1.000	*6.000	5.000	1.000
Boston	50	Dead	6.000	7.000	6.000	0.000
		Alive	11.000	24.000	4.000	0.000
	60	Dead	8.000	20.000	3.000	2.000
		Alive	18.000	58.000	10.000	3.000
	70	Dead	9.000	18.000	3.000	0.000
		Alive	15.000	26.000	1.000	1.000
Glamorgn	50	Dead	16.000	7.000	3.000	0.000
		Alive	16.000	20.000	8.000	1.000
	60	Dead	14.000	12.000	3.000	0.000
		Alive	27.000	39.000	10.000	4.000
	70	Dead	3.000	7.000	3.000	0.000
		Alive	12.000	11.000	4.000	1.000

\* indicates structural zero cells

```
Pearson Chi-square : 46.842 df : 47 p-value : 0.479
LR Chi-square      : 44.881 df : 47 p-value : 0.561
Raftery's BIC      : -260.538
Dissimilarity       : 10.168
```

#### Tests for Hierarchical Terms

Term Tested Hierarchically	The Model without the Term				Removal of Term from Model			
	ln(MLE)	Chi-square	df	p-value	Chi-square	df	p-value	
AGE	-190.460	132.866	53	0.000	87.984	6	0.000	
SURVIVES\$	-206.152	164.249	50	0.000	119.368	3	0.000	
TUMORS\$	-326.389	404.724	56	0.000	359.843	9	0.000	
CENTER\$	-177.829	107.604	61	0.000	62.722	14	0.000	
CENTER\$*AGE	-158.900	69.746	51	0.042	24.865	4	0.000	
CENTER\$*SURVIVES\$	-149.166	50.277	49	0.423	5.396	2	0.067	
CENTER\$*TUMORS\$	-162.289	76.522	53	0.019	31.641	6	0.000	

The model has a nonsignificant test of fit and so does a model without the *CENTER\$* by *SURVIVAL\$* interaction ( $p\text{-value} = 0.4226$ ).

### Eliminating Only the Young Women

Two of the extreme cells are from the youngest age group. What happens to the *CENTER\$* by *SURVIVE\$* effect if only these cells are defined as structural zeros? *HTERM* remains in effect.

The input, to declare these cells as structural zeros, is:

```
MODEL center$*age*survive$*tumor$ =,
      (age + survive$ + tumor$) # center$
ZERO CELL=1 1 2 2 CELL=1 1 2 3
ESTIMATE / DELTA=0.5
```

The following are the selections of the output:

```
Number of Cells (product of levels) : 72
Number of structural zero cells      : 2
Total count                          : 671

Pearson Chi-square : 50.261 df : 49 p-value : 0.423
LR Chi-square      : 49.115 df : 49 p-value : 0.469
Raftery's BIC      : -269.814
Dissimilarity       : 10.637
```

#### Tests for Hierarchical Terms

Term Tested Hierarchically	The Model without the Term ln (MLE)	Chi-square	df	p-value	Removal of Term from Model Chi-square	df	p-value
AGE	-221.256	188.370	55	0.000	139.254	6	0.000
SURVIVE\$	-210.369	166.596	52	0.000	117.481	3	0.000
TUMOR\$	-331.132	408.121	58	0.000	359.005	9	0.000
CENTER\$	-192.179	130.215	63	0.000	81.100	14	0.000
CENTER\$*AGE	-172.356	90.570	53	0.001	41.455	4	0.000
CENTER\$*SURVIVE\$	-153.888	53.633	51	0.374	4.517	2	0.104
CENTER\$*TUMOR\$	-169.047	83.952	55	0.007	34.837	6	0.000

When the two cells for the young women from Tokyo are excluded from the model estimation, the *CENTER\$* by *SURVIVE\$* effect is not needed ( $p\text{-value} = 0.3737$ ).

### Eliminating the Older Women

Here you define the two cells for the Tokyo women from the oldest age group as structural zeros.



The input is:

```
MODEL center$*age*survive$*tumor$ =,
      (age + survive$ + tumor$) # center$
ZERO CELL=1 3 2 1 CELL=1 3 2 2
ESTIMATE / DELTA=0.5
```

The following are the selections of the output:

Case frequencies determined by value of variable NUMBER

```
Number of Cells (product of levels) : 72
Number of structural zero cells      : 2
Total count                          : 757
```

```
Pearson Chi-square : 53.435 df : 49 p-value : 0.308
LR Chi-square      : 50.982 df : 49 p-value : 0.396
Raftery's BIC      : -273.856
Dissimilarity       : 9.458
```

#### Tests for Hierarchical Terms

Term Tested Hierarchically	The Model without the Term			Removal of Term from Model		
	ln(MLE)	Chi-square	df	p-value	Chi-square	df p-value
AGE	-203.305	147.406	55	0.000	96.423	6 0.000
SURVIVE\$	-238.968	218.731	52	0.000	167.749	3 0.000
TUMOR\$	-358.521	457.838	58	0.000	406.855	9 0.000
CENTER\$	-209.549	159.893	63	0.000	108.911	14 0.000
CENTER\$*AGE	-177.799	96.393	53	0.000	45.410	4 0.000
CENTER\$*SURVIVE\$	-161.382	63.560	51	0.111	12.577	2 0.002
CENTER\$*TUMOR\$	-171.123	83.041	55	0.009	32.058	6 0.000

When the two cells for the women from the older age group are treated as structural zeros, the case for removing the *CENTER\$* by *SURVIVE\$* effect is much weaker than when the cells for the younger women are structural zeros. Here, the inclusion of the effect results in a significant improvement in the fit of the model ( $p\text{-value} = 0.0019$ ).

## Conclusion

The structural zero feature allowed you to quickly focus on 2 of the 72 cells in your multiway table: the survivors under 50 from Tokyo, especially those with benign tumors with minimal inflammation. The overall survival rate for the 764 women is 72.5%, that for Tokyo is 79.3%, and that for the most unusual cell is 90.67%. Half of the Tokyo women under age 50 have *MinBenign* tumors (75 out of 151) and almost 10% of the 764 women (spread across 72 cells) are concentrated here. Possibly the protocol for study entry (including definition of a “tumor”) was executed differently at this center than at the others.



### Example 4

#### Tables without Analyses

If you want only a frequency table and no analysis, use TABULATE. Simply specify the table factors in the same order in which you want to view them from left to right. In other words, the last variable defines the columns of the table and cross-classifications of the preceding variables the rows.

For this example, we use data in the *CANCER* file. Here we use LOGLIN to display counts for a 3 by 3 by 2 by 4 table (72 cells) in two dozen lines.

The input is:

```
USE CANCER
LOGLIN
  FREQ number
  LABEL age / 50='Under 50', 60='50 to 69', 70='70 & Over'
  ORDER center$ / SORT=NONE
  ORDER tumor$ / SORT = 'MinBengn', 'MaxBengn', 'MinMalig', 'MaxMalig'
  TABULATE age * center$ * survive$ * tumor$
```

The output is:

Case frequencies determined by value of variable NUMBER

Number of Cells (product of levels) : 72  
Total count : 764

AGE	CENTER\$	SURVIVE\$	TUMOR\$			
			MinBengn	MaxBengn	MinMalig	MaxMalig
Under 50	Tokyo	Alive	68.000	9.000	26.000	25.000
		Dead	7.000	3.000	9.000	4.000
	Boston	Alive	24.000	0.000	11.000	4.000
		Dead	7.000	0.000	6.000	6.000
	Glamorgn	Alive	20.000	1.000	16.000	8.000
		Dead	7.000	0.000	16.000	3.000
50 to 69	Tokyo	Alive	46.000	5.000	20.000	18.000
		Dead	9.000	2.000	9.000	11.000
	Boston	Alive	58.000	3.000	18.000	10.000
		Dead	20.000	2.000	8.000	3.000
	Glamorgn	Alive	39.000	4.000	27.000	10.000
		Dead	12.000	0.000	14.000	3.000
70 & Over	Tokyo	Alive	6.000	1.000	1.000	5.000
		Dead	3.000	0.000	2.000	1.000
	Boston	Alive	26.000	1.000	15.000	1.000
		Dead	18.000	0.000	9.000	3.000
	Glamorgn	Alive	11.000	1.000	12.000	4.000
		Dead	7.000	0.000	3.000	3.000

## Computation

### Algorithms

Loglinear modeling implements the algorithms of Haberman (1973).

### References

- Agresti, A. (1984). *Analysis of ordinal categorical data*. New York: Wiley-Interscience.
- \* Agresti, A. (2002). *Categorical data analysis*, 2nd ed. New York: Wiley-Interscience.
- Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1977). *Discrete multivariate analysis: Theory and practice*. Cambridge, Mass: The MIT Press.
- Fienberg, S. E. (1980). *The analysis of cross-classified categorical data*, 2nd ed. Cambridge, Mass: MIT Press.
- Goodman, L.A. (1971). The analysis of multidimensional contingency tables: Stepwise procedures and direct estimation methods for building models for multiple classifications. *Technometrics*, 13, 33-61.
- Goodman, L. A. (1978). *Analyzing qualitative/categorical data: Loglinear models and latent structure analysis*. Cambridge, Mass: Abt Books.
- Haberman, S. J. (1973). Loglinear fit for contingency tables, algorithm AS 51. *Applied Statistics*, 21, 218-224.
- Haberman, S. J. (1978). *Analysis of qualitative data, Vol. 1: Introductory topics*. New York: Academic Press.
- Knock, D. and Burke, P. J. (1980). *Loglinear models*. Newbury Park: Sage.
- \* Morrison, D. F. (2004). *Multivariate statistical methods*, 4th ed. Pacific Grove, CA: Duxbury Press.

(\* indicates additional reference.)

# *Missing Value Analysis*

*Rick Marcantonio and Michael Pechnyo*

Missing value analysis helps address several concerns caused by incomplete data. Cases with missing values that are systematically different from cases without missing values can obscure the results. Also, missing data may reduce the precision of calculated statistics because there is less information than originally planned. Another concern is that the assumptions behind many statistical procedures are based on complete cases, and missing values can complicate the theory required.

The MISSING module displays and analyzes missing value patterns in data. The procedure computes maximum likelihood estimates of correlation, covariance, and cross-products of deviations matrices using either linear regression or an EM algorithm. You can downweight outliers using a normal or a t distribution.

Statistics computed include missing value patterns, means, correlations, variances and covariances, cross-products of deviations, and a pairwise frequency table. In addition, for EM estimation, SYSTAT reports Little's MCAR test. The correlation, covariance, or SSP matrix can be saved to a data file for further analyses. Alternatively, you can save imputed estimates in place of missing values.

Resampling procedures are available in this feature.

## *Statistical Background*

Even in the best designed and monitored study, observations can be missing—a subject inadvertently skips a question, a blood sample is ruined, or the recording equipment malfunctions. Because many classical statistical analyses require complete cases (no missing values), when data are incomplete it may be hard “to get off the ground.” That is, if the analyst wants to explore a new data set by, say, using a factor



analysis to identify redundant variables or sets of related variables, a cluster analysis to check for distinct subpopulations, or a stepwise discriminant analysis to see which variables differ among subgroups, there may be too few complete cases for an analysis. Alternatively, the complete cases may not fully represent the total sample, leading to biased results.

Analysis of missing values focuses on three issues:

- **Description of patterns.** How many missing values are there? Where are they located (specific cases and/or variables)? Are values missing randomly? For each variable, the word *pattern* indicates the dichotomized version of the variable—that is, a binary distribution where each value is *missing* or *present*. Also, when the same variables are missing for several cases, cases are said to have the same *pattern*.
- **Estimation of parameters, including means, covariances, and correlations.** Statistics are computed using either the EM (expectation maximization) algorithm or linear regression.
- **Imputation of values.** EM and regression methods are provided for estimating replacement values for the missing data.

Often it is necessary to run the MISSING procedure several times. You should:

- First, see the extent and pattern of missing values, and determine if values are missing randomly. At this point, you may want to delete cases and variables with large numbers of missing data and, most importantly, screen variables with skewed distributions for symmetrizing transformations before proceeding to the estimation or imputation phases.
- Next, study various estimates of descriptive statistics, possibly making a side step to check relations graphically when differences in estimates are found.
- Finally, impute values (estimate replacement values) and use graphics to assess the suitability of the filled-in values.

The use of a data matrix with imputed values may not be acceptable for a final report of results, but by using the approaches and methods described here, you may be able to find a subset of variables with enough complete cases for a meaningful analysis. You may omit variables simply because a large proportion of their values are missing; or, by making exploratory runs using the imputed data matrix, you may learn that some variables are redundant or have little relation to the outcome variables of interest. For example:



- In a stepwise regression, you may find that some variables have no relation to your outcome variable. Try rerunning the analysis with a smaller subset of candidate variables that has many more complete cases.
- In a factor analysis, you may identify one or more redundant variables. You might also learn this by examining an estimate of the correlation matrix in the MISSING procedure.

## ***Techniques for Handling Missing Values***

Over the years, many software users approached the missing data problem by using a pairwise complete method to compute a covariance or correlation matrix and then using this matrix as input for, say, a factor analysis. However, such a matrix may have eigenvalues less than 0, and some correlations may be computed from substantially different subsets of the cases. Other analysts use EM (expectation-maximization) or regression methods to estimate statistics or to impute data. Simulation studies indicate that pairwise estimates are often more distorted than estimates obtained via the EM method. In most algorithms, they are simply the first iteration of the EM method. A few analysts use multiple imputation, a computationally complex method that is not commonly available.

### ***Deletion Methods***

The two most common deletion methods are listwise and pairwise deletion. In listwise deletion, the analysis uses complete cases only. That is, the procedure removes from computations any observation with a value missing on any variable included in the analysis.

Pairwise deletion is listwise deletion done separately for every pair of selected variables. In other words, counts, sums of squares, and sums of cross-products are computed separately for every pair of variables in the file. With pairwise deletion, you get the same correlation (covariance, etc.) for two variables containing missing data if you select them alone or with other variables containing missing data. With listwise deletion, correlations under these two circumstances may differ, depending on the pattern of missing data among the other variables in the file.

Because it makes better use of the data than listwise deletion, pairwise deletion is a popular method for computing correlations on matrices with missing data. Many regression programs include it as a standard method for computing regression estimates from a covariance or correlation matrix.

Ironically, pairwise deletion is one of the worst ways to handle missing values. If as few as 20% of the values in a data matrix are missing, it is not difficult to find two correlations that were computed using substantially different subsets of the cases. In such cases, it is common to encounter error messages that the matrix is singular in regression programs and to get eigenvalues less than 0 in factor analysis.

But, more importantly, *classical statistical analyses require complete cases*. For exploration, this restriction can be circumvented by identifying one or more variables that are not needed, deleting them, and requesting the desired analysis—there should be more complete cases for this smaller set of variables.

If you have missing values, you may want to compare results from pairwise deletion with those from the EM method. Or, you may want to take the time to replace the missing values in the raw data by examining similar cases or variables with nonmissing values.

### ***Imputation Methods***

Deletion methods attempt to restrict computations to complete cases by eliminating cases or variables that are incomplete. Imputation methods, on the other hand, replace missing data with hypothesized values, resulting in a “complete” data set consisting of observed and imputed values. Analyses that require complete cases can then be applied to the resulting data.

#### **Unconditional Mean Imputation**

One common imputation technique replaces all missing values for a variable with the mean of the observed values for that variable. Although it is highly unlikely that the missing values, if actually observed, would all lie at the center of the distribution for the variable, the most likely value for each missing point is the mean. Placing all missing values at the center of the distribution, however, underestimates the variances and covariances for the variables.

Let’s look at a simple case. Consider two variables, X and Y, having a positive correlation. X has a mean of 5 and a variance of 1. Y has a mean of 13.5 and a variance

of 3.25. The covariance between X and Y equals 1.80. The data in the X and Y columns of the following table represent ten observations on these variables.

Case	X	Y	X'	Y'
1	4.65	13.85	4.67	13.86
2	6.21	16.41	6.21	15.22*
3	6.63	15.68	6.64	15.68
4	4.94	15.76	4.95	15.77
5	7.21	17.70	4.98*	17.70
6	5.09	13.44	5.09	15.22*
7	6.08	15.64	4.98*	15.64
8	4.19	12.94	4.20	12.95
9	3.09	10.67	3.09	15.22*
10	5.19	14.95	4.98*	14.96
Mean	5.33	14.71	4.98	15.22
Variance	1.51	4.06	.95	1.55
Covariance	2.29		.33	

Suppose that the Y values for cases 2, 6, and 9 and the X values for 5, 7, and 10 could not be observed. Simple mean imputation yields the data in columns X' and Y' (imputed values are marked with an asterisk). Notice:

- For X' and Y', the mean for the ten cases equals the mean for the seven observed cases.
- The variances for X' and Y' underestimate the corresponding true variances.
- The covariance between X' and Y' underestimates the true covariance between X and Y.

The systematic underestimation of the variances and covariances suggests that any conclusions drawn from analyses using the imputed data are suspect.

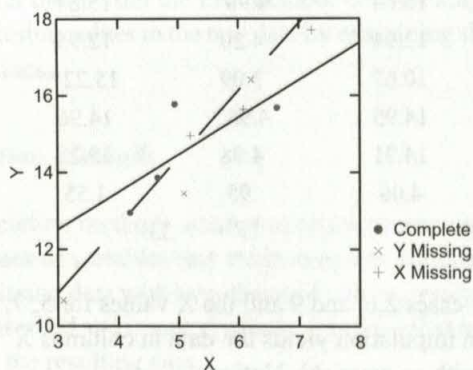
### Regression Imputation

Buck (1960) suggested an alternative procedure for imputation using conditional means. In Buck's method, the sample means and covariance matrix for the complete cases are used as estimates for the corresponding population parameters. These estimates are subsequently used to compute linear regressions of the variables with



missing values on the variables without missing values *for each case*. The resulting regression equations allow you to predict the missing values from the observed values.

The following plot illustrates the technique for the ten cases presented above. Cases with missing Y values could be placed at any Y value for the corresponding observed X value; cases with missing X values could be placed at any X value for the corresponding observed Y value. In this display, we place missing values at points corresponding to the complete sample (if we had been able to observe it). The solid line represents the regression of Y on X and should be used to impute values for cases lacking Y values. The dashed line indicates the regression of X on Y and is used to impute values when the X value is missing.



The two regression lines result in the following imputed estimates appearing in columns X'' and Y'':



Case	X'	Y'	X''	Y''
1	4.67	13.86	4.67	13.86
2	6.21	15.22*	6.21	15.64*
3	6.64	15.68	6.64	15.68
4	4.95	15.77	4.95	15.77
5	4.98*	17.70	6.90*	17.70
6	5.09	15.22*	5.09	14.55*
7	4.98	15.64	5.72*	15.64
8	4.20	12.95	4.20	12.95
9	3.09	15.22*	3.09	12.60*
10	4.98*	14.96	5.33*	14.96
Mean	4.98	15.22	5.27	14.93
Variance	.95	1.55	1.34	2.29
Covariance	.33		1.57	

Compare the mean, variance, and covariance estimates with those obtained using unconditional mean imputation (columns X' and Y'). The variance for Y and the covariance still underestimate the true values, but to a lesser extent than found previously.

### Other Imputation Methods

Replacing missing values by means (unconditional or conditional) is one approach to imputation. Other techniques found in the literature include:

- replacing missing data with values selected randomly from a distribution for each missing value.
- replacing missing data with values selected from cases not included in the analysis.
- adding a random residual to the conditional mean estimates.
- imputating multiple values for each missing item.

None of these methods, however, should be used as a panacea for solving the missing data problem. For a complete discussion of these methods, see Little and Rubin (2002).

**EM Method**

Instead of pairwise deletion, many data analysts prefer to use an EM algorithm when estimating correlations, covariances, or an SSCP matrix. EM uses the maximum likelihood method to compute the estimates. This procedure defines a model for the partially missing data and bases inferences on the likelihood under that model. Each iteration consists of an E step and an M step. The E step finds the conditional expectation of the log likelihood based on complete data, with respect to the missing data model, given the observed values and current estimates of the parameters. For the M step, maximum likelihood estimation is performed for this expectation. "Missing" is enclosed in quotation marks because the missing values are not being directly filled but, rather, functions of them are used in the log-likelihood. Estimation iterates between these two steps until the parameters converge.

Returning to the previous data set, the EM imputed values appear in the final two columns of the following table:

Case	X''	Y''	X'''	Y'''
1	4.67	13.86	4.67	13.86
2	6.21	15.64*	6.21	16.00*
3	6.64	15.68	6.64	15.68
4	4.95	15.77	4.95	15.77
5	6.90*	17.70	6.86*	17.70
6	5.09	14.55*	5.09	14.86*
7	5.72*	15.64	5.62*	15.64
8	4.20	12.95	4.20	12.95
9	3.09	12.60*	3.09	12.83*
10	5.33*	14.96	5.21*	14.96
Mean	5.27	14.93	5.25	15.02
Variance	1.34	2.29	1.51	2.55
Covariance	1.57		1.54	

For this simple example, the regression and EM results are very similar. However, when data are missing for several variables across cases, the EM method generally outperforms regression imputation. The latter technique cannot capture covariances between jointly missing data, nor does it lead to maximum likelihood estimates based on observed data.

If you compute the covariance matrix for the imputed data, the estimates will differ from the variances shown above. The EM algorithm estimates two sets of parameters (the means and covariances) with corresponding sufficient statistics (the sums of values, and the sums of cross-products). In the M step, the first set of statistics yields the EM mean estimates and the second set yields the EM covariance estimates. Using the imputed data to estimate the covariances and variances ignores any relationships between the presence or absence of data across variables. In effect, one set of sufficient statistics is being used to estimate both sets of parameters. As a result, the variances estimated from the imputed data always underestimate the variances produced by the EM algorithm. See Little and Rubin for details.

By default for the EM method, the Missing Value procedure assumes that the data follow a normal distribution. If you know that the tails of the distributions are longer than those of a normal distribution, you can request that a  $t$  distribution with  $n$  degrees of freedom be used in constructing the likelihood function ( $n$  is specified by the user). A second option also provides a distribution with longer tails. You specify the ratio of standard deviations of a mixed normal distribution and the mixture proportion of the two distributions. This assumes that only the standard deviations of the distributions differ, not the means.

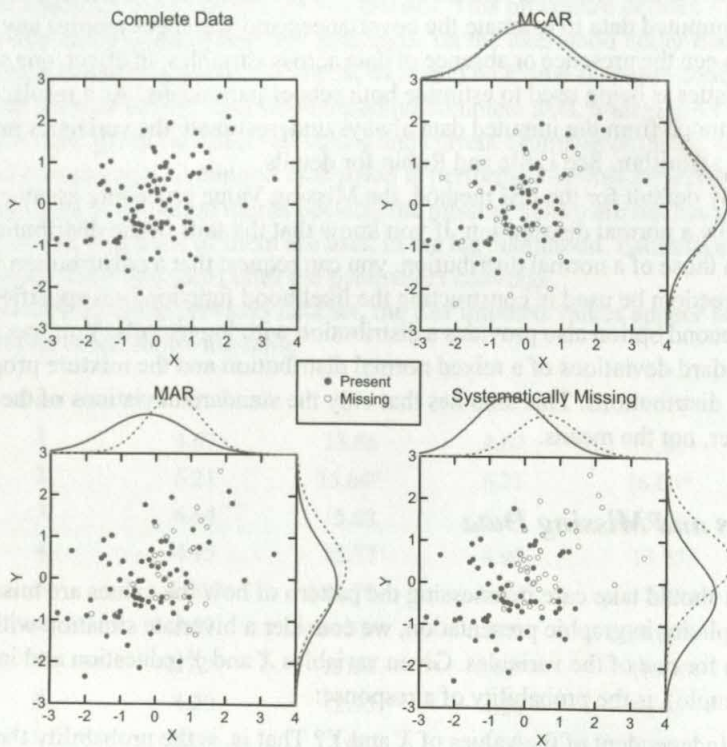
## ***Randomness and Missing Data***

You should take care in assessing the pattern of how the values are missing. For simplicity in graphic presentation, we consider a bivariate situation with incomplete data for one of the variables. Given variables  $X$  and  $Y$  (education and income, for example), is the probability of a response:

- Independent of the values of  $X$  and  $Y$ ? That is, is the probability that income is recorded the same for all people regardless of their education or incomes? The recorded or observed values of income form a random subsample of the true incomes for all of the people in the sample. Little and Rubin call this pattern MCAR (Missing Completely At Random).
- Dependent on  $X$  but not on  $Y$ ? In this case, the probability that income is recorded depends on the subject's education, so the probability varies by education but not by income *within that education group*. This pattern is called MAR (Missing At Random).
- Dependent on  $Y$  and possibly  $X$  also? In this case, the probability that income is present varies by the value of income within each education group. This is not an unusual pattern for real-world applications.



The following figure illustrates these missing data situations. In the upper left plot, the data contain no missing values. The remaining three plots depict the relationship between  $X$  and  $Y$  when approximately 30% of the data are missing. The border plots display the approximate distribution of cases for each situation.



In the MCAR plot, notice the random scatter of missing and present data. Missing observations occur for both low and high values of both variables. The distribution of the missing values is indistinguishable from the distribution of observed values for both variables. If data follow this pattern, the pairwise deletion, EM, and regression methods give consistent and unbiased estimates of correlations and covariances.

In the MAR plot, the missing values tend to occur for large values of  $X$ . However, the unobserved values are spread throughout the range of  $Y$ . The distributions for the missing and complete groups are practically identical when focusing on  $Y$ . In other words, the probability of nonresponse is independent of  $Y$ . However, two distributions emerge along the  $X$  variable. The missing value distribution (shown with a dashed



line) shifts toward higher values. The probability of observing nonresponse increases as  $X$  increases.

The pairwise, EM, and regression methods may still provide good estimates if the data are missing at random. For example, in a study of education and income, the subjects with low education may have more missing income values. If education is MCAR and if, for a given level of education, income is MCAR, pairwise, EM, and regression methods may still yield good estimates.

If the data are MAR and the assumption that the distributions are normal, mixed normal, or  $t$  with specific degrees of freedom is met, the EM method yields maximum likelihood estimates of means, standard deviations, covariances, and correlations. Be sure to check the data for outliers and to determine whether symmetrizing transformations are required before applying the technique, however.

In the final plot, the missing values appear in the upper right area of the plot. In contrast to the MAR plot, the value of  $Y$  influences the probability of nonresponse; the higher the  $Y$  value, the more likely the value will be missing. The distributions along both axes have much less overlap, with unique centers appearing for each group of cases. This situation is not an unusual pattern for real-world applications, but no current estimation methods are appropriate for data of this type.

### ***Testing for Randomness***

The Little (1988) chi-square statistic for testing whether values are missing completely at random is printed with EM matrices. The test computes the Mahalanobis distance between parameter estimates based on listwise complete data and parameter estimates resulting from the EM algorithm. The resulting sum is referred to a chi-square distribution with degrees of freedom based on the number of patterns of missing data in the data set. If the test is rejected, the EM and listwise estimates are sufficiently "far" enough apart to warrant further examination, and certainly tells one that analysis based on listwise estimates MAY be biased.

Another method for testing for randomness involves dividing a variable into two groups based on whether data are missing or present for another variable. The means for the two groups can be compared using a  $t$ -statistic; if the values are not missing randomly, the test statistic will be large. However, be aware that while a sizable  $t$  statistic does indicate a departure from randomness, a small  $t$  may be no confirmation that values are missing randomly. Sadly, there is no magic test for MAR.

## ***A Final Caution***

Imputed data are not complete. Although missing values do not occur in imputed data, imputation does not replace them with values that would have been observed had all data been available. If you use imputed data in analyses, you should control for the imputation. For example, suppose you use the EM estimates in a regression, the degrees of freedom for the error term should be adjusted back down to either the listwise complete value or some other reasonable estimate.

To us, none of the approaches to estimation and imputation should be viewed as a magic black box. While the EM and regression methods allow a specific way in which the values of one variable may be related to another, a good data analyst will want to ferret out possible problems in how the data are sampled, recorded, or otherwise fail to conform to the study protocol—for example, which regions of a multivariate space are sparse because data are missing? It is hard to separate the selection of an appropriate method for estimation or imputation from the basic data screening process.

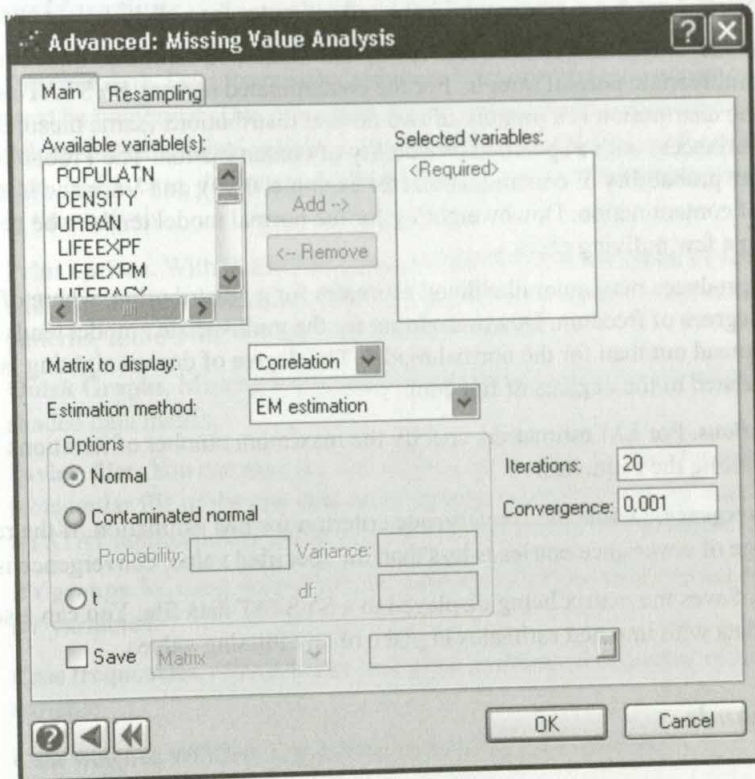
## ***Missing Value Analysis in SYSTAT***

### ***Missing Value Analysis Dialog Box***

To analyze missing values, from the menus choose:

Advanced

Missing Value Analysis...



SYSTAT treats all selected variables as continuous (numeric) data. Select a matrix to compute and a method for handling missing data.

**Matrix to display.** SYSTAT computes the correlation, covariance, or SSCP matrix.

**Estimation method.** Two estimation methods are available:

- **EM estimation.** Requests the EM algorithm to estimate Pearson correlation, covariance, or SSCP matrices. Little's MCAR test is shown with a display of the pattern of missing values.
- **Regression substitution.** Uses multiple linear regression to impute estimates for missing values. For each case, SYSTAT uses linear regression on the observed variables to predict values for the missing variables.

You can downweight outliers using a Normal, contaminated normal, or t distribution. The following options are available:



- Normal produces maximum likelihood estimates for a multivariate normal sample.
- Contaminated normal produces maximum likelihood estimates for a contaminated multivariate normal sample. For the contaminated normal, SYSTAT assumes that the distribution is a mixture of two normal distributions (same mean, different variances) with a specified probability of contamination. The Probability value is the probability of contamination (for example, 0.10), and Variance is the variance of contamination. Downweighting for the normal model tends to be concentrated in a few outlying cases.
- $t$  produces maximum likelihood estimates for a  $t$  distribution, where  $df$  is the degrees of freedom. Downweighting for the multivariate  $t$  model tends to be more spread out than for the normal model. The degree of downweighting is inversely related to the degrees of freedom.

**Iterations.** For EM estimation, specify the maximum number of iterations for computing the estimates.

**Convergence.** Define the convergence criterion for EM estimation. If the relative change of covariance entries is less than the specified value, convergence is assumed.

**Save.** Saves the matrix being displayed to a SYSTAT data file. You can also save the raw data with imputed estimates in place of any missing values.

## Using Commands

Select your data by typing `USE filename`. Continue with:

```
MISSING
MODEL varlist
SAVE outfile / DATA
ESTIMATE / MATRIX = CORRELATION
                    COVARIANCE
                    SSCP,
NORMAL = n1, n2,
T = df,
ITER = n,
CONV = n,
REGRESSION
BOOT = SAMPLE(m, n) SIMPLE(m, n) JACK
```

Omitting the DATA option from SAVE results in the current matrix being saved to *outfile*.



## Usage Considerations

**Types of data.** Data for missing value analysis must be rectangular and all variables must be numerical. This procedure should not be used to estimate missing categorical values, but categorical variables can be used to estimate values for missing continuous data. In this case, dummy code the categories and use the resulting indicator variables in the analysis.

**Print options.** With PLENGTH LONG, SYSTAT prints the mean of each variable. In addition, for EM estimation, SYSTAT prints an iteration history, missing value patterns, Little's MCAR test, and mean estimates.

**Quick Graphs.** Missing value analysis produces a cases-by-variables plot similar to a shaded data matrix.

**Saving files.** You can save the correlation, covariance, or SSCP matrix, or save a rectangular file of the raw data with missing values replaced by imputed estimates. SYSTAT automatically defines the type of file as CORR, COVA, SSCP, or RECT.

**BY groups.** Missing value analysis produces separate analyses for each level of any BY variables.

**Case frequencies.** FREQUENCY <variable> increases the number of cases by the FREQ variable.

**Case weights.** WEIGHT is available in missing value analysis.

## Examples

### Example 1

#### Missing Values: Preliminary Examinations

Where are the missing values located? How extensive are they? If a value is missing for one variable, does it tend to be missing for one or more other variables? Conversely, if a value is present for one variable, do values tend to be missing for other specific variables? Is the pattern of missing values related to values of another variable?

You may need to uncover patterns of incomplete data in order to:

- select enough complete cases for a meaningful analysis. If you omit a few variables, or even just one, does the sample size of complete cases increase dramatically?

- select a method of estimation or imputation. If, for example, you plan to use complete cases for a final analysis, you need to verify that values are missing *completely* at random, missing at random, or missing nonrandomly.
- understand how results may be biased or distorted because of a failure to meet necessary assumptions about randomness of the missing values.

In this example, we explore the *WORLD95m* data for patterns of how values are missing. We focus on descriptive statistics to explore variable distributions and reveal the amount of missing data.

The input is:

```
USE WORLD95M
CSTATISTICS POPULATN DENSITY URBAN LIFEEXPF LIFEEXPM,
  LITERACY POP INCR BABYMORT GDP CAP CALORIES,
  BIRTH_RT DEATH_RT B_TO_D FERTILTY LIT MALE,
  LIT_FEMA / MEAN MEDIAN SD SES SKEWNESS N
```

The output is:

	POPULATN	DENSITY	URBAN	LIFEEXPF
N of Cases	109.000	109.000	108.000	109.000
Median	10400.000	64.000	60.000	74.000
Arithmetic Mean	47723.881	203.415	56.528	70.156
Standard Deviation	146726.364	675.705	24.203	10.572
Skewness(G1)	6.592	6.887	-0.308	-1.109
Standard Error of Skewness	0.231	0.231	0.233	0.231
	LIFEEXPM			
N of Cases	109.000			
Median	67.000			
Arithmetic Mean	64.917			
Standard Deviation	9.273			
Skewness(G1)	-1.080			
Standard Error of Skewness	0.231			
	LITERACY	POP_INCR	BABYMORT	GDP_CAP
N of Cases	107.000	109.000	109.000	109.000
Median	88.000	1.800	27.700	2995.000
Arithmetic Mean	78.336	1.682	42.313	5859.982
Standard Deviation	22.883	1.198	38.079	6479.836
Skewness(G1)	-0.994	0.324	1.090	1.146
Standard Error of Skewness	0.234	0.231	0.231	0.231
	CALORIES			
N of Cases	75.000			
Median	2653.000			
Arithmetic Mean	2753.827			
Standard Deviation	567.828			
Skewness(G1)	0.170			
Standard Error of Skewness	0.277			

	BIRTH_RT	DEATH_RT	B_TO_D	FERTILTY
N of Cases	109.000	108.000	108.000	107.000
Median	25.000	9.000	2.667	3.050
Arithmetic Mean	25.923	9.557	3.204	3.563
Standard Deviation	12.361	4.253	2.125	1.902
Skewness(G1)	0.446	1.308	1.829	0.664
Standard Error of Skewness	0.231	0.233	0.233	0.234
	LIT_MALE			
N of Cases	85.000			
Median	87.000			
Arithmetic Mean	78.729			
Standard Deviation	20.445			
Skewness(G1)	-0.851			
Standard Error of Skewness	0.261			
	LIT_FEMA			
N of Cases	85.000			
Median	71.000			
Arithmetic Mean	67.259			
Standard Deviation	28.607			
Skewness(G1)	-0.504			
Standard Error of Skewness	0.261			

This output provides your first look, variable by variable, at the extent of incomplete data. Because means and standard deviations are computed using all available data for each variable, the sample sizes vary from variable to variable. The total number of observations is 109. The number of values present for each variable is reported as 'N of cases'. For calories, 75 countries (cases) report a value, so  $109 - 75$ , or 34, do not. That is, calories is missing for  $34 / 109 = 31.2\%$  of the cases. The female and male literacy rates (*lit\_fema* and *lit\_male*) are each missing for 22% of the cases. Eight variables have no missing values, and five others have from 0.9% to 1.8% missing values.

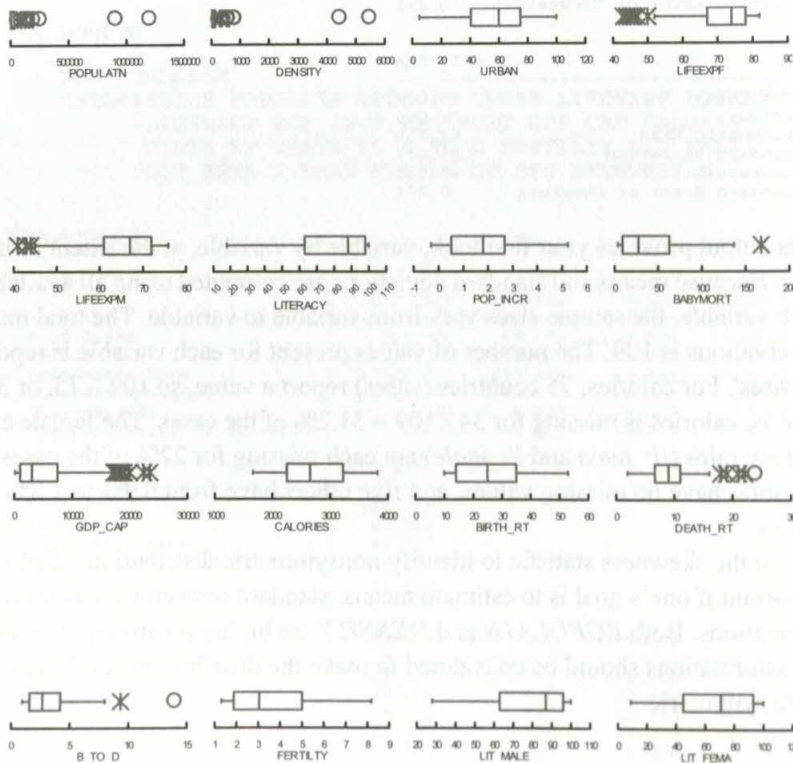
Use the skewness statistic to identify nonsymmetric distributions. Symmetry is important if one's goal is to estimate means, standard deviations, covariances, or correlations. Both *POPULATN* and *DENSITY* are highly positively skewed. Transformations should be considered to make the distributions of these variables more symmetric.



## Boxplots and Transformations

Boxplots and stem-and-leaf plots provide a visual display of distributions and assist in identifying outliers. To generate boxplots for the *WORLD95m* data, the input is:

```
USE WORLD95M
DENSITY POPULATN DENSITY URBAN LIFEEXPF LIFEEXPM,
LITERACY POP_INCR BABYMORT GDP_CAP CALORIES,
BIRTH_RT DEATH_RT B_TO_D FERTILTY LIT_MALE,
LIT_FEMA / BOX
```



*POPULATN*, *DENSITY*, *GDP\_CAP* and *DEATH\_RT* all contain many extreme cases and outliers. Transforming these variables may eliminate these problematic cases and improve the symmetry of the distributions.

The log transformation improves the distributions of these variables considerably. Here we plot the boxplots for the original data next to the boxplots for the log-

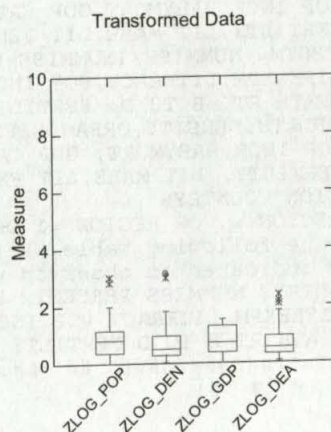
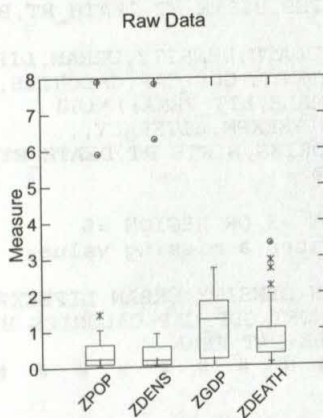


transformed data. In order to display the four distributions within each plot, we standardize the variables before plotting.

```

USE WORLD95M
LET ZPOP = POPULATN
LET ZDENS = DENSITY
LET ZGDP = GDP_CAP
LET ZDEATH = DEATH_RT
LET ZLOG_POP = L10(POPULATN)
LET ZLOG_DEN = L10(DENSITY)
LET ZLOG_GDP = L10(GDP_CAP)
LET ZLOG_DEA = L10(DEATH_RT)
STANDARDIZE ZPOP ZDENS ZGDP ZDEATH,
            ZLOG_POP ZLOG_DEN ZLOG_GDP ZLOG_DEA
BEGIN
DENSITY ZPOP ZDENS ZGDP ZDEATH / REPEAT BOX XLAB='',
                                TITLE='Raw Data' LOC=-3IN,0IN
DENSITY ZLOG_POP ZLOG_DEN ZLOG_GDP ZLOG_DEA / REPEAT BOX,
                                XLAB='' TITLE='Transformed Data',
                                YMIN=-5 YMAX=10 LOC=3IN,0IN
END 'Boxplots'

```



For each variable, the number of extreme cases decreases after applying the transformation. In addition, cases identified as extreme occur at both ends of the distribution for the transformed data. In contrast, extreme cases for the raw data correspond only to the high end of the distributions. The improvement in the distributions suggests transforming these variables to logarithms before applying any missing value analysis.

## Example 2

### Casewise Pattern Table

A casewise pattern table is a picture of the data file that highlights the location of missing observations. Each column in the display represents the values of a variable; each row represents the data for one case. This display is used to see if particular cases and/or variables have too little complete data to use and also to see if variables (or groups of variables) have values missing nonrandomly.

In this example, we create this layout using the MIS function. In addition, we recode the variables as (0,1) indicator variables, in which a 1 indicates a missing value and a 0 indicates an observed value. To save space, the Eastern European, African, and Latin American countries are omitted.

The input is:

```
USE WORLD95M
LET NUMMISS=MIS (POPULATN,DENSITY,URBAN,LIFEEXPF,LIFEEXPM,LITERACY,,
  POP_INCR,BABYMORT,GDP_CAP,CALORIES,BIRTH_RT,DEATH_RT,B_TO_D,,
  FERTILTY,LIT_MALE,LIT_FEMA)
LET PERCENTM= NUMMISS/(NUMMISS+NUM(POPULATN,DENSITY,URBAN,LIFEEXPF,,
  LIFEEXPM,LITERACY,POP_INCR, BABYMORT, GDP_CAP,CALORIES,BIRTH_RT,,
  DEATH_RT, B_TO_D, FERTILTY,LIT_MALE,LIT_FEMA))*100
LET (POPULATN,DENSITY,URBAN,LIFEEXPF,LIFEEXPM,LITERACY,,
  POP_INCR,BABYMORT, GDP_CAP, CALORIES,BIRTH_RT,DEATH_RT,B_TO_D,,
  FERTILTY, LIT_MALE,LIT_FEMA) = @ = .
SORT REGION COUNTRY$
SELECT REGION =. OR REGION =1 OR REGION =3 OR REGION =5
REM 'In the following table, a 1 indicates a missing value.'
REM 'A 0 indicates an observed value.'
LIST COUNTRY$ NUMMISS PERCENTM POPULATN DENSITY URBAN LIFEEXPF,
  LIFEEXPM LITERACY POP_INCR BABYMORT GDP_CAP CALORIES BIRTH_RT,
  DEATH_RT B_TO_D FERTILTY LIT_MALE LIT_FEMA
!!/FORMAT='##### ## ##.## | | # # # # # # # # # # # #
# # # #'
```

Because USA and Canada have missing values for *REGION2*, we select cases where *REGION2* is missing to include these countries in the table. We also sort the cases by geographical region and by country name, yielding an alphabetical listing of countries within each region.

The output is:

Data for the following results were selected according to  
SELECT REGION =. OR REGION =1 OR REGION =3 OR REGION =5

Case	COUNTRY\$ LIFEEXPF CALORIES LIT_FEMA	NUMMISS LIFEEXPM BIRTH_RT	PERCENTM LITERACY DEATH_RT	POPULATN POP_INCR B_TO_D	DENSITY BABYMORT FERTILTY
1	Australia 80.00000 3216.00000 100.00000	0.00000 74.00000 15.00000	0.00000 100.00000 8.00000	17800.00000 1.38000 1.87500	2.30000 7.30000 1.90000
2	Austria 79.00000 3495.00000	2.00000 73.00000 12.00000	12.50000 99.00000 11.00000	8000.00000 0.20000 1.09091	94.00000 6.70000 1.50000
3	Belgium 79.00000 .	3.00000 73.00000 12.00000	18.75000 99.00000 11.00000	10100.00000 0.20000 1.09091	329.00000 7.20000 1.70000
4	Canada 81.00000 3482.00000	2.00000 74.00000 14.00000	12.50000 97.00000 8.00000	29100.00000 0.70000 1.75000	2.80000 6.80000 1.80000
5	Denmark 79.00000 3628.00000	2.00000 73.00000 12.00000	12.50000 99.00000 12.00000	5200.00000 0.10000 1.00000	120.00000 6.60000 1.70000
6	Finland 80.00000 3253.00000	2.00000 72.00000 13.00000	12.50000 100.00000 10.00000	5100.00000 0.30000 1.30000	39.00000 5.30000 1.80000
7	France 82.00000 3465.00000	2.00000 74.00000 13.00000	12.50000 99.00000 9.30000	58000.00000 0.47000 1.39785	105.00000 6.70000 1.80000
8	Germany 79.00000 3443.00000	2.00000 73.00000 11.00000	12.50000 99.00000 11.00000	81200.00000 0.36000 1.00000	227.00000 6.50000 1.47000
9	Greece 80.00000 3825.00000 89.00000	0.00000 75.00000 10.00000	0.00000 93.00000 10.00000	10400.00000 0.84000 1.00000	80.00000 8.20000 1.50000
10	Iceland 81.00000 .	3.00000 76.00000 16.00000	18.75000 100.00000 7.00000	263.00000 1.10000 2.28571	2.50000 4.00000 2.11000
11	Ireland 78.00000 3778.00000	2.00000 73.00000 14.00000	12.50000 98.00000 9.00000	3600.00000 0.30000 1.55556	51.00000 7.40000 1.99000
12	Italy 81.00000 3504.00000 96.00000	0.00000 74.00000 11.00000	0.00000 97.00000 10.00000	58100.00000 0.21000 1.10000	188.00000 7.60000 1.30000
13	Netherlands 81.00000 3151.00000	2.00000 75.00000 13.00000	12.50000 99.00000 9.00000	15400.00000 0.58000 1.44444	366.00000 6.30000 1.58000
14	New Zealand 80.00000 3362.00000	2.00000 73.00000 16.00000	12.50000 99.00000 8.00000	3524.00000 0.57000 2.00000	13.00000 8.90000 2.03000
15	Norway 81.00000 3326.00000	2.00000 74.00000 13.00000	12.50000 99.00000 10.00000	4300.00000 0.40000 1.30000	11.00000 6.30000 2.00000



16	Portugal	1.00000	6.25000	10500.00000	108.00000
	78.00000	71.00000	85.00000	0.36000	9.20000
	.	12.00000	10.00000	1.20000	1.50000
17	82.00000	.	.	.	.
	Spain	0.00000	0.00000	39200.00000	77.00000
	81.00000	74.00000	95.00000	0.25000	6.90000
18	3572.00000	11.00000	9.00000	1.22222	1.40000
	93.00000	.	.	.	.
	Sweden	2.00000	12.50000	8800.00000	19.00000
19	81.00000	75.00000	99.00000	0.52000	5.70000
	2960.00000	14.00000	11.00000	1.27273	2.10000
	.	.	.	.	.
20	Switzerland	2.00000	12.50000	7000.00000	170.00000
	82.00000	75.00000	99.00000	0.70000	6.20000
	3562.00000	12.00000	9.00000	1.33333	1.60000
21	UK	2.00000	12.50000	58400.00000	237.00000
	80.00000	74.00000	99.00000	0.20000	7.20000
	3149.00000	13.00000	11.00000	1.18182	1.83000
22	USA	0.00000	0.00000	260800.00000	26.00000
	79.00000	73.00000	97.00000	0.99000	8.11000
	3671.00000	15.00000	9.00000	1.66667	2.06000
23	97.00000	.	.	.	.
	Afghanistan	1.00000	6.25000	20500.00000	25.00000
	44.00000	45.00000	29.00000	2.80000	168.00000
24	.	53.00000	22.00000	2.40909	6.90000
	14.00000	.	.	.	.
	Bangladesh	0.00000	0.00000	125000.00000	800.00000
25	53.00000	53.00000	35.00000	2.40000	106.00000
	2021.00000	35.00000	11.00000	3.18182	4.70000
	22.00000	.	.	.	.
26	Cambodia	0.00000	0.00000	10000.00000	55.00000
	52.00000	50.00000	35.00000	2.90000	112.00000
	2166.00000	45.00000	16.00000	2.81250	5.81000
27	22.00000	.	.	.	.
	China	0.00000	0.00000	1.20520E+006	124.00000
	69.00000	67.00000	78.00000	1.10000	52.00000
28	2639.00000	21.00000	7.00000	3.00000	1.84000
	68.00000	.	.	.	.
	.	.	.	.	.

Case URBAN  
GDP CAP  
LIT\_MALE

1	85.00000
	16848.00000
	100.00000
2	58.00000
	18396.00000
	.
3	96.00000
	17912.00000
	.
4	77.00000
	19904.00000
	.
5	85.00000
	18277.00000
	.
6	60.00000
	15877.00000
	.



7	73.00000 18944.00000 .
8	85.00000 17539.00000 .
9	63.00000 8060.00000 98.00000
10	91.00000 17241.00000 .
11	57.00000 12170.00000 .
12	69.00000 17500.00000 98.00000
13	89.00000 17245.00000 .
14	84.00000 14381.00000 .
15	75.00000 17755.00000 .
16	34.00000 9000.00000 89.00000
17	78.00000 13047.00000 97.00000
18	84.00000 16900.00000 .
19	62.00000 22384.00000 .
20	89.00000 15974.00000 .
21	75.00000 23474.00000 97.00000
36	18.00000 205.00000 44.00000
37	16.00000 202.00000 47.00000
38	12.00000 260.00000 48.00000

39      26.00000  
          377.00000  
          87.00000

Case	COUNTRY\$ LIFEEXPF CALORIES LIT_FEMA	NUMMISS LIFEEXPM BIRTH_RT	PERCENTM LITERACY DEATH_RT	POPULATN POP_INCR B_TO_D	DENSITY BABYMORT FERTILTY
40	Hong Kong 80.00000 . 64.00000	1.00000 75.00000 13.00000	6.25000 77.00000 6.00000	5800.00000 -0.09000 2.16667	5494.00000 5.80000 1.40000
41	India 59.00000 2229.00000 39.00000	0.00000 58.00000 29.00000	0.00000 52.00000 10.00000	911600.00000 1.90000 2.90000	283.00000 79.00000 4.48000
42	Indonesia 65.00000 2750.00000 68.00000	0.00000 61.00000 24.00000	0.00000 77.00000 9.00000	199700.00000 1.60000 2.66667	102.00000 68.00000 2.80000
43	Japan 82.00000 2956.00000 .	2.00000 76.00000 11.00000	12.50000 99.00000 7.00000	125500.00000 0.30000 1.57143	330.00000 4.40000 1.55000
44	Malaysia 72.00000 2774.00000 70.00000	0.00000 66.00000 29.00000	0.00000 78.00000 5.00000	19500.00000 2.30000 5.80000	58.00000 25.60000 3.51000
45	N. Korea 73.00000 . 99.00000	1.00000 67.00000 24.00000	6.25000 99.00000 5.50000	23100.00000 1.83000 4.36364	189.00000 27.70000 2.40000
46	Pakistan 58.00000 . 21.00000	1.00000 57.00000 42.00000	6.25000 35.00000 10.00000	128100.00000 2.80000 4.20000	143.00000 101.00000 6.43000
47	Philippines 68.00000 2375.00000 90.00000	0.00000 63.00000 27.00000	0.00000 90.00000 7.00000	69800.00000 1.92000 3.85714	221.00000 51.00000 3.35000
48	S. Korea 74.00000 . 99.00000	1.00000 68.00000 16.00000	6.25000 96.00000 6.00000	45000.00000 1.00000 2.66667	447.00000 21.70000 1.65000
49	Singapore 79.00000 3198.00000 84.00000	0.00000 73.00000 16.00000	0.00000 88.00000 6.00000	2900.00000 1.20000 2.66667	4456.00000 5.70000 1.88000
50	Taiwan 78.00000 . .	6.00000 72.00000 15.60000	37.50000 91.00000 .	20944.00000 0.92000 .	582.00000 5.10000 .
51	Thailand 72.00000 2316.00000 90.00000	0.00000 65.00000 19.00000	0.00000 93.00000 6.00000	59400.00000 1.40000 3.16667	115.00000 37.00000 2.10000
52	Vietnam 68.00000 2233.00000 83.00000	0.00000 63.00000 27.00000	0.00000 88.00000 8.00000	73100.00000 1.78000 3.37500	218.00000 46.00000 3.33000
72	Armenia 75.00000 . 100.00000	1.00000 68.00000 23.00000	6.25000 98.00000 6.00000	3700.00000 1.40000 3.83333	126.00000 27.00000 3.19000
73	Azerbaijan 75.00000 . 100.00000	1.00000 67.00000 23.00000	6.25000 98.00000 7.00000	7400.00000 1.40000 3.28571	86.00000 35.00000 2.80000

## Missing Value Analysis

74	Bahrain	1.00000	6.25000	600.00000	828.00000
	74.00000	71.00000	77.00000	2.40000	25.00000
	.	29.00000	4.00000	7.25000	3.96000
	55.00000				
75	Egypt	0.00000	0.00000	60000.00000	57.00000
	63.00000	60.00000	48.00000	1.95000	76.40000
	3336.00000	29.00000	9.00000	3.22222	3.77000
	34.00000				
76	Iran	0.00000	0.00000	65600.00000	39.00000
	67.00000	65.00000	54.00000	3.46000	60.00000
	3181.00000	42.00000	8.00000	5.25000	6.33000
	43.00000				
77	Iraq	0.00000	0.00000	19900.00000	44.00000
	68.00000	65.00000	60.00000	3.70000	67.00000
	2887.00000	44.00000	7.00000	6.28571	6.71000
	49.00000				
78	Israel	1.00000	6.25000	5400.00000	238.00000
	80.00000	76.00000	92.00000	2.22000	8.60000
	.	21.00000	7.00000	3.00000	2.83000
	89.00000				
79	Jordan	0.00000	0.00000	3961.00000	42.00000
	74.00000	70.00000	80.00000	3.30000	34.00000
	2634.00000	39.00000	5.00000	7.80000	5.64000
	70.00000				
80	Kuwait	0.00000	0.00000	1800.00000	97.00000
	78.00000	73.00000	73.00000	5.24000	12.50000
	3195.00000	28.00000	2.00000	14.00000	4.00000
	67.00000				
81	Lebanon	1.00000	6.25000	3620.00000	343.00000
	71.00000	67.00000	80.00000	2.00000	39.50000
	.	27.00000	7.00000	3.85714	3.39000
	73.00000				
82	Libya	0.00000	0.00000	5500.00000	2.80000
	65.00000	62.00000	64.00000	3.70000	63.00000
	3324.00000	45.00000	8.00000	5.62500	6.40000
	50.00000				
83	Oman	4.00000	25.00000	1900.00000	7.80000
	70.00000	66.00000	.	3.46000	36.70000
	.	40.00000	5.00000	8.00000	6.53000
	.				

Case URBAN  
GDP\_CAP  
LIT\_MALE

40	94.00000
	14641.00000
	90.00000
41	26.00000
	275.00000
	64.00000
42	29.00000
	681.00000
	84.00000
43	77.00000
	19860.00000
	.
44	43.00000
	2995.00000
	86.00000
45	60.00000
	1000.00000
	99.00000
46	32.00000

	406.00000	00000.0	00000.0	00000.0
	47.00000	00000.0	00000.0	00000.0
47	43.00000	00000.0	00000.0	00000.0
	867.00000	00000.0	00000.0	00000.0
	90.00000	00000.0	00000.0	00000.0
48	72.00000	00000.0	00000.0	00000.0
	6627.00000	00000.0	00000.0	00000.0
	99.00000	00000.0	00000.0	00000.0
49	100.00000	00000.0	00000.0	00000.0
	14990.00000	00000.0	00000.0	00000.0
	93.00000	00000.0	00000.0	00000.0
50	71.00000	00000.0	00000.0	00000.0
	7055.00000	00000.0	00000.0	00000.0
51	22.00000	00000.0	00000.0	00000.0
	1800.00000	00000.0	00000.0	00000.0
	96.00000	00000.0	00000.0	00000.0
52	20.00000	00000.0	00000.0	00000.0
	230.00000	00000.0	00000.0	00000.0
	93.00000	00000.0	00000.0	00000.0
72	68.00000	00000.0	00000.0	00000.0
	5000.00000	00000.0	00000.0	00000.0
	100.00000	00000.0	00000.0	00000.0
73	54.00000	00000.0	00000.0	00000.0
	3000.00000	00000.0	00000.0	00000.0
	100.00000	00000.0	00000.0	00000.0
74	83.00000	00000.0	00000.0	00000.0
	7875.00000	00000.0	00000.0	00000.0
	55.00000	00000.0	00000.0	00000.0
75	44.00000	00000.0	00000.0	00000.0
	748.00000	00000.0	00000.0	00000.0
	63.00000	00000.0	00000.0	00000.0
76	57.00000	00000.0	00000.0	00000.0
	1500.00000	00000.0	00000.0	00000.0
	64.00000	00000.0	00000.0	00000.0
77	72.00000	00000.0	00000.0	00000.0
	1955.00000	00000.0	00000.0	00000.0
	70.00000	00000.0	00000.0	00000.0
78	92.00000	00000.0	00000.0	00000.0
	13066.00000	00000.0	00000.0	00000.0
	95.00000	00000.0	00000.0	00000.0
79	68.00000	00000.0	00000.0	00000.0
	1157.00000	00000.0	00000.0	00000.0
	89.00000	00000.0	00000.0	00000.0
80	96.00000	00000.0	00000.0	00000.0
	6818.00000	00000.0	00000.0	00000.0
	77.00000	00000.0	00000.0	00000.0
81	84.00000	00000.0	00000.0	00000.0
	1429.00000	00000.0	00000.0	00000.0
	88.00000	00000.0	00000.0	00000.0
82	82.00000	00000.0	00000.0	00000.0
	5910.00000	00000.0	00000.0	00000.0
	75.00000	00000.0	00000.0	00000.0



## Missing Value Analysis

83 11.00000  
7467.00000  
.

Case	COUNTRY\$ LIFEEXPF CALORIES LIT_FEMA	NUMMISS LIFEEXPM BIRTH_RT	PERCENTM LITERACY DEATH_RT	POPULATN POP_INCR B_TO_D	DENSITY BABYMORT FERTILTY
84	Saudi Arabia	0.00000	0.00000	18000.00000	7.70000
	70.00000	66.00000	62.00000	3.20000	52.00000
	2874.00000	38.00000	6.00000	6.33333	6.67000
	48.00000				
85	Syria	1.00000	6.25000	14900.00000	74.00000
	68.00000	65.00000	64.00000	3.70000	43.00000
	.	44.00000	6.00000	7.33333	6.65000
	51.00000				
86	Turkey	0.00000	0.00000	62200.00000	79.00000
	73.00000	69.00000	81.00000	2.02000	49.00000
	3236.00000	26.00000	6.00000	4.33333	3.21000
	71.00000				
87	U.Arab Em.	1.00000	6.25000	2800.00000	32.00000
	74.00000	70.00000	68.00000	4.80000	22.00000
	.	28.00000	3.00000	9.33333	4.50000
	63.00000				
88	Uzbekistan	1.00000	6.25000	22600.00000	50.00000
	72.00000	65.00000	97.00000	2.13000	53.00000
	.	30.00000	7.00000	4.28571	3.73000
	100.00000				

Case URBAN  
GDP\_CAP  
LIT\_MALE

84	77.00000
	6651.00000
	73.00000
85	50.00000
	2436.00000
	78.00000
86	61.00000
	3721.00000
	90.00000
87	81.00000
	14193.00000
	70.00000
88	41.00000
	1350.00000
	100.00000

The 1's show that when female literacy is missing, male literacy is missing too (see the final two columns). *LIT\_MALE* and *LIT\_FEMA* are missing frequently for European countries, but calories is missing more often for Middle Eastern countries. In the complete sample, 37.5% of Taiwan's data are missing, 25% of Oman's data are missing, and so forth.

**Sorted Pattern Table**

In a sorted pattern table, cases and variables are sorted by the patterns of the missing data. Complete cases are not included.

The input is:

```

USE WORLD95M
LET NUMMISS=MIS (POPULATN,DENSITY,URBAN,LIFEEXPF,LIFEEXPM,,
LITERACY, POP_INCR, BABYMORT, GDP_CAP,,
CALORIES,BIRTH_RT,DEATH_RT, B_TO_D, FERTILTY,,
LIT_MALE,LIT_FEMA)
LET PERCENTM = NUMMISS/(NUMMISS+NUM (POPULATN,DENSITY,URBAN,,
LIFEEXPF, LIFEEXPM, LITERACY, POP_INCR, BABYMORT,GDP_CAP,,
CALORIES, BIRTH_RT, DEATH_RT, B_TO_D, FERTILTY,LIT_MALE,,
LIT_FEMA))*100
DSAVE WORLD95N
TRANPOSE POPULATN DENSITY URBAN LIFEEXPF LIFEEXPM LITERACY,
POP_INCR BABYMORT GDP_CAP CALORIES, BIRTH_RT DEATH_RT B_TO_D,
FERTILTY LIT_MALE LIT_FEMA NUMMISS PERCENTM
LET NUMMISS=MIS (COL(1)..COL(109))
SORT NUMMISS
TRANPOSE
DSAVE RECODE

MERGE WORLD95N (COUNTRY$ NUMMISS PERCENTM) RECODE
DROP LABEL$

LET (POPULATN DENSITY URBAN LIFEEXPF LIFEEXPM LITERACY,
POP_INCR BABYMORT GDP_CAP CALORIES BIRTH_RT DEATH_RT B_TO_D,
FERTILTY LIT_MALE,LIT_FEMA) = @ = .
SORT NUMMISS_WORLD95N
SELECT NUMMISS_WORLD95N > 1
REM 'In the following table, a 1 indicates a missing value.'
REM 'A 0 indicates an observed value.'
REM LIST / FORMAT='##### ## ##.## || # # # # # # # #
# # # # # #'
LIST COUNTRY$ NUMMISS_WORLD95N PERCENTM_WORLD95N POPULATN,
DENSITY LIFEEXPF LIFEEXPM POP_INCR BABYMORT GDP_CAP BIRTH_RT,
NUMMISS_RECODE PERCENTM_RECODE URBAN DEATH_RT B_TO_D,
LITERACY FERTILTY LIT_MALE LIT_FEMA CALORIES
!!/FORMAT='##### ## ##.## || # # # # # # # # # #
# # # # # #'

```

To shorten the output, we omit countries with one missing value. *CALORIES* is missing for most of the omitted cases.

Data for the following results were selected according to  
SELECT NUMMISS\_WORLD95N > 1

Case	COUNTRY\$ LIFEEXPM PERCENTM_REC LIT_MALE	NUMMISS_WORLD9- 5N POP_INCR URBAN LIT_FEMA	PERCENTM_WORLD- 95N BABYMORT DEATH_RT CALORIES	POPULATN GDP_CAP B_TO_D
87	Austria 73.000 12.500 .	2.000 0.200 58.000 .	12.500 6.700 11.000 3495.000	8000.000 18396.000 1.091 29100.000
88	Canada 74.000 12.500 .	2.000 0.700 77.000 .	12.500 6.800 8.000 3482.000	19904.000 1.750 5200.000
89	Denmark 73.000 12.500 .	2.000 0.100 85.000 .	12.500 6.600 12.000 3628.000	18277.000 1.000 5100.000
90	Finland 72.000 12.500 .	2.000 0.300 60.000 .	12.500 5.300 10.000 3253.000	15877.000 1.300 58000.000
91	France 74.000 12.500 .	2.000 0.470 73.000 .	12.500 6.700 9.300 3465.000	18944.000 1.398 81200.000
92	Germany 73.000 12.500 .	2.000 0.360 85.000 .	12.500 6.500 11.000 3443.000	17539.000 1.000 3600.000
93	Ireland 73.000 12.500 .	2.000 0.300 57.000 .	12.500 7.400 9.000 3778.000	12170.000 1.556 125500.000
94	Japan 76.000 12.500 .	2.000 0.300 77.000 .	12.500 4.400 7.000 2956.000	19860.000 1.571 15400.000
95	Netherlands 75.000 12.500 .	2.000 0.580 89.000 .	12.500 6.300 9.000 3151.000	17245.000 1.444 3524.000
96	New Zealand 73.000 12.500 .	2.000 0.570 84.000 .	12.500 8.900 8.000 3362.000	14381.000 2.000 4300.000
97	Norway 74.000 12.500 .	2.000 0.400 75.000 .	12.500 6.300 10.000 3326.000	17755.000 1.300 23400.000
98	Romania 69.000 12.500 .	2.000 0.060 54.000 .	12.500 20.300 10.000 3155.000	2702.000 1.400 8800.000
99	Sweden 75.000 12.500 .	2.000 0.520 84.000 .	12.500 5.700 11.000 2960.000	16900.000 1.273 7000.000
100	Switzerland 75.000 12.500 .	2.000 0.700 62.000 .	12.500 6.200 9.000 3562.000	22384.000 1.333 .

101	UK	2.000	12.500	58400.000
	74.000	0.200	7.200	15974.000
	12.500	89.000	11.000	1.182
	.	.	3149.000	.
102	Belgium	3.000	18.750	10100.000
	73.000	0.200	7.200	17912.000
	18.750	96.000	11.000	1.091
	.	.	.	.
103	Bulgaria	3.000	18.750	8900.000
	69.000	-0.200	12.000	3831.000
	18.750	68.000	12.000	1.083
	.	.	.	.
104	Croatia	3.000	18.750	4900.000
	70.000	-0.100	8.700	5487.000
	18.750	51.000	11.000	1.000
	.	.	.	.
105	Iceland	3.000	18.750	263.000
	76.000	1.100	4.000	17241.000
	18.750	91.000	7.000	2.286
	.	.	.	.
106	South Africa	3.000	18.750	43900.000
	62.000	2.600	47.100	3128.000
	18.750	49.000	8.000	4.250
	.	.	.	.
107	Bosnia	4.000	25.000	4600.000
	72.000	0.700	12.700	3098.000
	25.000	36.000	6.390	2.191
	.	.	.	.
108	Czech Rep.	4.000	25.000	10400.000
	69.000	0.210	9.300	7311.000
	25.000	.	11.100	1.171
	.	.	3632.000	.
109	Oman	4.000	25.000	1900.000
	66.000	3.460	36.700	7467.000
	25.000	11.000	5.000	8.000
	.	.	.	.
110	Taiwan	6.000	37.500	20944.000
	72.000	0.920	5.100	7055.000
	37.500	71.000	.	.
	.	.	.	.

Case	DENSITY BIRTH RT LITERACY	LIFEEXPF NUMMISS RECODE FERTILTY
------	---------------------------------	--

87	94.000	79.000
	12.000	2.000
	99.000	1.500
88	2.800	81.000
	14.000	2.000
	97.000	1.800
89	120.000	79.000
	12.000	2.000
	99.000	1.700
90	39.000	80.000
	13.000	2.000
	100.000	1.800
91	105.000	82.000
	13.000	2.000
	99.000	1.800
92	227.000	79.000
	11.000	2.000
	99.000	1.470



## Missing Value Analysis

93	51.000	78.000
	14.000	2.000
	98.000	1.990
94	330.000	82.000
	11.000	2.000
	99.000	1.550
95	366.000	81.000
	13.000	2.000
	99.000	1.580
96	13.000	80.000
	16.000	2.000
	99.000	2.030
97	11.000	81.000
	13.000	2.000
	99.000	2.000
98	96.000	75.000
	14.000	2.000
	96.000	1.820
99	19.000	81.000
	14.000	2.000
	99.000	2.100
100	170.000	82.000
	12.000	2.000
	99.000	1.600
101	237.000	80.000
	13.000	2.000
	99.000	1.830
102	329.000	79.000
	12.000	3.000
	99.000	1.700
103	79.000	75.000
	13.000	3.000
	93.000	1.800
104	85.000	77.000
	11.000	3.000
	97.000	1.650
105	2.500	81.000
	16.000	3.000
	100.000	2.110
106	35.000	68.000
	34.000	3.000
	76.000	4.370
107	87.000	78.000
	14.000	4.000
	86.000	.
108	132.000	77.000
	13.000	4.000

	.	1.840
109	7.800	70.000
	40.000	4.000
	.	6.530
110	582.000	78.000
	15.600	6.000
	91.000	.

The last three columns are *LIT\_MALE*, *LIT\_FEMA*, and *CALORIES*, and the four last cases are Oman, Bosnia, Czech Rep., and Taiwan, because they have the most values missing. Recalling that cases with one missing value are not included and that this missing value is usually *CALORIES*, it is easy to see that when *CALORIES* is missing, the literacy rates for females and males tend to be present. For larger data files, the most common patterns may be less apparent.

### Example 3 Correlation Estimation

In this example, we continue to use the *WORLD95m* data used in the "Preliminary Examinations" example, now requesting estimates of correlations. Even though we established that values are nonrandomly missing, we request listwise estimates so that they can be compared later with estimates obtained by the pairwise, and EM.

The input is:

```
USE WORLD95M
LET LOG_DEA = L10(DEATH_RT)
FORMAT 6,3
CORR
NOTE 'Listwise Deletion'
PEARSON LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
        BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
        FERTILTY URBAN LITERACY LIT_FEMA,
        LIT_MALE CALORIES / LISTWISE
```

The output is:

Listwise Deletion

Number of Observations: 59

#### Means

LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP	BIRTH_RT
4.237	1.660	65.831	61.339	2.214	57.729	3.129	31.492

## Missing Value Analysis

## Means (contd...)

LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA	LIT_MALE	CALORIES
0.945	3.776	4.303	49.763	69.576	62.119	75.356	2.589E+003

## Pearson Correlation Matrix

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP
LOG_POP	1.000						
LOG_DEN	0.282	1.000					
LIFEEXPF	0.038	0.004	1.000				
LIFEEXPM	0.059	0.023	0.987	1.000			
POP_INCR	-0.299	-0.206	-0.392	-0.325	1.000		
BABYMORT	-0.009	-0.037	-0.951	-0.931	0.420	1.000	
LOG_GDP	-0.139	-0.216	0.766	0.736	-0.363	-0.745	1.000
BIRTH_RT	-0.223	-0.136	-0.817	-0.773	0.776	0.809	-0.674
LOG_DEA	0.029	-0.015	-0.801	-0.823	-0.102	0.742	-0.478
B_TO_D	-0.269	-0.072	0.270	0.318	0.692	-0.231	0.083
FERTILTY	-0.240	-0.142	-0.790	-0.747	0.755	0.784	-0.586
URBAN	-0.141	-0.226	0.741	0.717	-0.192	-0.705	0.786
LITERACY	0.082	0.004	0.827	0.785	-0.567	-0.891	0.642
LIT_FEMA	0.109	0.072	0.815	0.773	-0.580	-0.856	0.602
LIT_MALE	0.176	0.097	0.754	0.727	-0.542	-0.805	0.580
CALORIES	0.142	-0.012	0.716	0.711	-0.393	-0.701	0.803

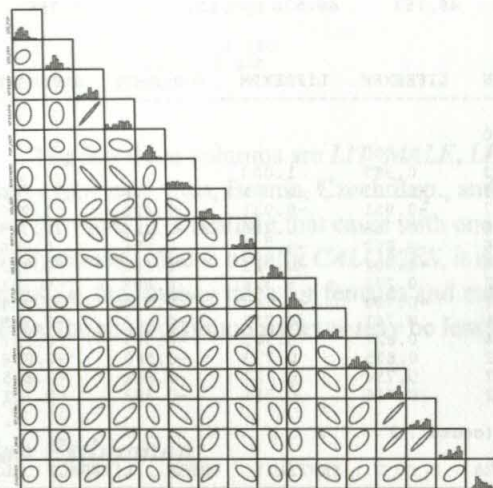
## Pearson Correlation Matrix (contd...)

	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA
LOG_POP							
LOG_DEN							
LIFEEXPF							
LIFEEXPM							
POP_INCR							
BABYMORT							
LOG_GDP							
BIRTH_RT	1.000						
LOG_DEA	0.468	1.000					
B_TO_D	0.188	-0.731	1.000				
FERTILTY	0.968	0.503	0.152	1.000			
URBAN	-0.566	-0.583	0.261	-0.533	1.000		
LITERACY	-0.822	-0.589	0.043	-0.814	0.614	1.000	
LIT_FEMA	-0.811	-0.570	0.032	-0.819	0.634	0.963	1.000
LIT_MALE	-0.756	-0.529	0.029	-0.759	0.595	0.939	0.960
CALORIES	-0.658	-0.407	0.040	-0.581	0.674	0.575	0.548

## Pearson Correlation Matrix (contd...)

	LIT_MALE	CALORIES
LOG_POP		
LOG_DEN		
LIFEEXPF		
LIFEEXPM		
POP_INCR		
BABYMORT		
LOG_GDP		
BIRTH_RT		
LOG_DEA		
B_TO_D		
FERTILTY		
URBAN		
LITERACY		
LIT_FEMA		
LIT_MALE	1.000	
CALORIES	0.576	1.000

## Scatter Plot Matrix



Of the 109 cases in the file, 50 have missing data. All statistics reported here are based on the remaining 59 cases. If you compute the means for these variables using CSTATISTICS, the values will differ. The latter procedure deletes cases on a variable-by-variable basis, instead of deleting a case if it has a missing value on any variable.

**Pairwise Deletion**

A table of frequency counts for each pair of variables provides a picture of the pattern of incomplete data. SYSTAT displays this table when using pairwise deletion in CORR or when using PLENGTH MEDIUM in MISSING.



The input is:

```
USE WORLD95M
LET LOG_DEA = L10 (DEATH_RT)
FORMAT 6,3
CORR
NOTE 'Pairwise Deletion'
PEARSON LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
        BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
        FERTILTY URBAN LITERACY LIT_FEMA,
        LIT_MALE CALORIES / PAIRWISE
```

The output is:

#### Pairwise Deletion

##### Means

LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP	BIRTH_RT
4.114	1.784	70.156	64.917	1.682	42.313	3.422	25.923

##### Means (contd...)

LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA	LIT_MALE	CALORIES
0.941	3.204	3.563	56.528	78.336	67.259	78.729	2.754E+003

##### Pearson Correlation Matrix

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP
LOG_POP	1.000						
LOG_DEN	0.143	1.000					
LIFEEXPF	-0.088	0.126	1.000				
LIFEEXPM	-0.082	0.153	0.982	1.000			
POP_INCR	-0.078	-0.252	-0.579	-0.502	1.000		
BABYMORT	0.109	-0.152	-0.962	-0.936	0.602	1.000	
LOG_GDP	-0.217	0.004	0.831	0.805	-0.557	-0.824	1.000
BIRTH_RT	-0.027	-0.216	-0.862	-0.805	0.861	0.865	-0.769
LOG_DEA	0.089	-0.064	-0.587	-0.640	-0.206	0.534	-0.322
B_TO_D	-0.153	-0.111	-0.087	-0.011	0.800	0.118	-0.209
FERTILTY	-0.060	-0.223	-0.838	-0.783	0.840	0.833	-0.693
URBAN	-0.138	0.015	0.743	0.730	-0.375	-0.718	0.754
LITERACY	-0.050	0.084	0.865	0.809	-0.699	-0.900	0.732
LIT_FEMA	0.005	0.113	0.819	0.745	-0.638	-0.843	0.632
LIT_MALE	0.076	0.138	0.777	0.717	-0.619	-0.809	0.611
CALORIES	0.046	0.050	0.775	0.765	-0.609	-0.777	0.847

##### Pearson Correlation Matrix (contd...)

	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA
LOG_POP							
LOG_DEN							
LIFEEXPF							
LIFEEXPM							
POP_INCR							
BABYMORT							
LOG_GDP							
BIRTH_RT	1.000						
LOG_DEA	0.230	1.000					
B_TO_D	0.483	-0.690	1.000				
FERTILTY	0.975	0.268	0.452	1.000			
URBAN	-0.629	-0.431	-0.032	-0.619	1.000		
LITERACY	-0.869	-0.385	-0.271	-0.866	0.650	1.000	
LIT_FEMA							1.000

LIT_FEMA	-0.835	-0.442	-0.148	-0.839	0.612	0.973	1.000
LIT_MALE	-0.794	-0.414	-0.153	-0.796	0.587	0.948	0.964
CALORIES	-0.762	-0.267	-0.240	-0.696	0.692	0.682	0.548

Pearson Correlation Matrix (contd...)

	LIT_MALE	CALORIES
LOG_POP		
LOG_DEN		
LIFEEXPF		
LIFEEXPM		
POP_INCR		
BABYMORT		
LOG_GDP		
BIRTH_RT		
LOG_DEA		
B_TO_D		
FERTILTY		
URBAN		
LITERACY		
LIT_FEMA	1.000	
LIT_MALE	0.576	1.000
CALORIES		

Pairwise Frequency Table

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP
LOG_POP	109						
LOG_DEN	109	109					
LIFEEXPF	109	109	109				
LIFEEXPM	109	109	109	109			
POP_INCR	109	109	109	109	109		
BABYMORT	109	109	109	109	109	109	
LOG_GDP	109	109	109	109	109	109	109
BIRTH_RT	109	109	109	109	109	109	109
LOG_DEA	108	108	108	108	108	108	108
B_TO_D	108	108	108	108	108	108	108
FERTILTY	107	107	107	107	107	107	107
URBAN	108	108	108	108	108	108	108
LITERACY	107	107	107	107	107	107	107
LIT_FEMA	85	85	85	85	85	85	85
LIT_MALE	85	85	85	85	85	85	85
CALORIES	75	75	75	75	75	75	75

Pairwise Frequency Table (contd...)

	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA
LOG_POP							
LOG_DEN							
LIFEEXPF							
LIFEEXPM							

Pairwise Frequency Table (contd...)

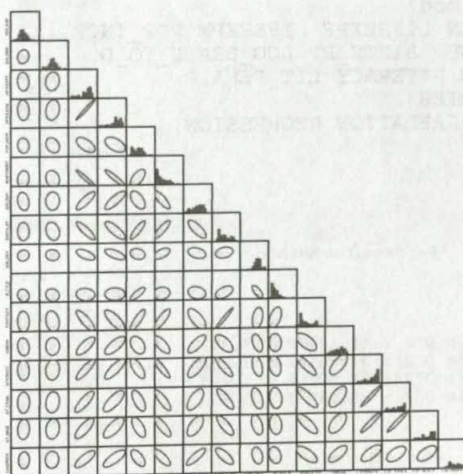
	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA
LOG_POP							
LOG_DEN							
LIFEEXPF							
LIFEEXPM							
POP_INCR							
BABYMORT							
LOG_GDP							
BIRTH_RT	109						
LOG_DEA	108	108					

B TO D	108	108	108			
FERTILTY	107	107	107	107		
URBAN	108	107	107	106	108	
LITERACY	107	106	106	105	107	107
LIT_FEMA	85	85	85	85	85	85
LIT_MALE	85	85	85	85	85	85
CALORIES	75	75	75	75	74	74

Pairwise Frequency Table (contd...)

	LIT_MALE	CALORIES
LOG_POP		
LOG_DEN		
LIFEEXPF		
LIFEEXPM		
POP_INCR		
BABYMORT		
LOG_GDP		
BIRTH_RT		
LOG_DEA		
B_TO_D		
FERTILTY		
URBAN		
LITERACY		
LIT_FEMA		
LIT_MALE	85	
CALORIES	59	75

## Scatter Plot Matrix



In contrast to listwise deletion, the number of cases used to compute each correlation and mean varies with the variable(s) involved. The mean computations use all observed cases for each variable. The correlation computations involve all cases that

have observed values for both variables. The pairwise frequency table displays the number of cases used to calculate each correlation.

The sample size for each variable is reported on the diagonal of the table; sample sizes for complete pairs of cases, off the diagonal. *CALORIES* alone has 75 values, but when paired with male or female literacy, the count of cases with both values drops to 59. If you need a set of variables for a multivariate analysis, it would be wise to omit *CALORIES* or the male and female literacy rates. Otherwise, if these variables are essential to your analysis, be concerned that results may be biased due to the fact they are not missing randomly.

### Regression Method

We now use the regression method for estimating the correlation matrix.

The input is:

```
USE WORLD95M
LET LOG_DEA = L10(DEATH_RT)
FORMAT 6,3
MISSING
NOTE 'Regression Method'
MODEL LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
      BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
      FERTILTY URBAN LITERACY LIT_FEMA,
      LIT_MALE CALORIES
ESTIMATE 7 MATRIX=CORRELATION REGRESSION
```

The output is:

#### Regression Method

Mahalanobis D<sup>2</sup> and z-score  
NOTE:

```
Case is an outlier. Mahalanobis D^2 : 41.670 Z : 3.307
Case is an outlier. Mahalanobis D^2 : 39.861 Z : 3.286
Case is an outlier. Mahalanobis D^2 : 68.621 Z : 5.419
Case is an outlier. Mahalanobis D^2 : 38.981 Z : 3.050
```

#### Missing Value Patterns

N of Cases	Missing Value Patterns (X=nonmissing; .=missing)
26	XXXXXXXXXXXXX-X.
59	XXXXXXXXXXXXX-XX
15	XXXXXXXXXXXXX.-



```

.X
5 XXXXXXXXXXXX.-
..
1 XXXXXXXXXXXX.XX.-
..
1 XXXXXXXXXXXX...-
.X
1 XXXXXXXXXXXX.-
..
1 XXXXXXXX...XX.-
..

```

## Regression Substitution Estimate of Means

LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP	BIRTH_RT
4.114	1.784	70.156	64.917	1.682	42.313	3.422	25.923

## Regression Substitution Estimate of Means (contd...)

LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA	LIT_MALE	CALORIES
0.941	3.201	3.533	56.614	78.440	69.423	80.383	2.769E+003

## Regression Substitution Estimated Correlation Matrix

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP
LOG_POP	1.000						
LOG_DEN	0.143	1.000					
LIFEEXPF	-0.088	0.126	1.000				
LIFEEXPM	-0.082	0.153	0.982	1.000			
POP_INCR	-0.078	-0.252	-0.579	-0.502	1.000		
BABYMORT	0.109	-0.152	-0.962	-0.936	0.602	1.000	
LOG_GDP	-0.217	0.004	0.831	0.805	-0.557	-0.824	1.000
BIRTH_RT	-0.027	-0.216	-0.862	-0.805	0.861	0.865	-0.769
LOG_DEA	0.087	-0.069	-0.587	-0.640	-0.203	0.535	-0.323
B_TO_D	-0.153	-0.112	-0.088	-0.012	0.799	0.119	-0.209
FERTILTY	-0.056	-0.232	-0.839	-0.785	0.841	0.835	-0.692
URBAN	-0.138	0.017	0.743	0.729	-0.376	-0.717	0.753
LITERACY	-0.046	0.092	0.864	0.807	-0.698	-0.899	0.728
LIT_FEMA	0.010	0.124	0.785	0.718	-0.611	-0.801	0.588
LIT_MALE	0.079	0.152	0.747	0.693	-0.597	-0.770	0.575
CALORIES	0.024	0.059	0.720	0.712	-0.529	-0.704	0.789

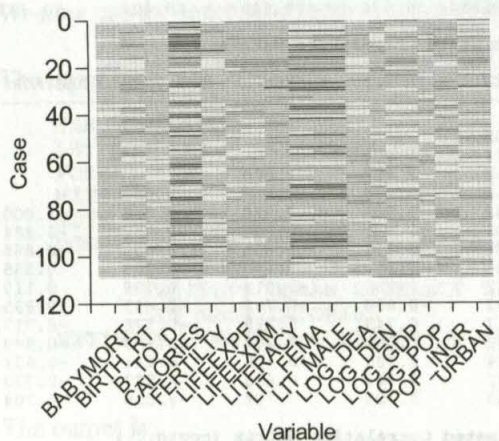
## Regression Substitution Estimated Correlation Matrix (contd...)

	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA
LOG_POP							
LOG_DEN							
LIFEEXPF							
LIFEEXPM							
POP_INCR							
BABYMORT							
LOG_GDP							
BIRTH_RT	1.000						
LOG_DEA	0.232	1.000					
B_TO_D	0.483	-0.689	1.000				
FERTILTY	0.975	0.274	0.454	1.000			
URBAN	-0.628	-0.428	-0.036	-0.607	1.000		
LITERACY	-0.867	-0.371	-0.277	-0.860	0.643	1.000	
LIT_FEMA	-0.782	-0.386	-0.216	-0.801	0.598	0.928	1.000
LIT_MALE	-0.747	-0.354	-0.223	-0.763	0.578	0.904	0.961
CALORIES	-0.682	-0.214	-0.232	-0.625	0.613	0.610	0.489

## Regression Substitution Estimated Correlation Matrix (contd...)

	LIT MALE	CALORIES
LOG POP		
LOG DEN		
LIFEEXPF		
LIFEEXPM		
POP INCR		
BABYMORT		
LOG GDP		
BIRTH RT		
LOG DEA		
B TO D		
FERTILT		
URBAN		
LITERACY		
LIT FEMA		
LIT MALE	1.000	
CALORIES	0.525	1.000

Missing Values Plot



Pairwise Frequency Table

	LOG POP	LOG DEN	LIFEEXPF	LIFEEXPM	POP INCR	BABYMORT
LOG POP	1.090E+002					
LOG DEN	1.090E+002	1.090E+002				
LIFEEXPF	1.090E+002	1.090E+002	1.090E+002			
LIFEEXPM	1.090E+002	1.090E+002	1.090E+002	1.090E+002		
POP INCR	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002	
BABYMORT	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002
LOG GDP	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002
BIRTH RT	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002
LOG DEA	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002
B TO D	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002
FERTILT	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002
URBAN	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002
LITERACY	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002
LIT FEMA	85.000	85.000	85.000	85.000	85.000	85.000
LIT MALE	85.000	85.000	85.000	85.000	85.000	85.000
CALORIES	75.000	75.000	75.000	75.000	75.000	75.000

Pairwise Frequency Table (contd...)

	LOG GDP	BIRTH RT	LOG DEA	B_TO_D	FERTILTY	URBAN
LOG POP						
LOG DEN						
LIFEEXPF						
LIFEEXPM						
POP INCR						
BABYMORT						
LOG GDP	1.090E+002					
BIRTH RT	1.090E+002	1.090E+002				
LOG DEA	1.080E+002	1.080E+002	1.080E+002			
B_TO_D	1.080E+002	1.080E+002	1.080E+002	1.080E+002		
FERTILTY	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002	
URBAN	1.080E+002	1.080E+002	1.070E+002	1.070E+002	1.060E+002	1.080E+002
LITERACY	1.070E+002	1.070E+002	1.060E+002	1.060E+002	1.050E+002	1.070E+002
LIT_FEMA	85.000	85.000	85.000	85.000	85.000	85.000
LIT_MALE	85.000	85.000	85.000	85.000	85.000	85.000
CALORIES	75.000	75.000	75.000	75.000	75.000	74.000

Pairwise Frequency Table (contd...)

	LITERACY	LIT_FEMA	LIT_MALE	CALORIES
LOG POP				
LOG DEN				
LIFEEXPF				
LIFEEXPM				
POP INCR				
BABYMORT				
LOG GDP				
BIRTH RT				
LOG DEA				
B_TO_D				
FERTILTY				
URBAN				
LITERACY	1.070E+002			
LIT_FEMA	85.000	85.000		
LIT_MALE	85.000	85.000	85.000	
CALORIES	74.000	59.000	59.000	75.000

In the Missing Value Patterns display, the patterns of missing values across variables are tabulated. An X indicates an observed value for a variable; a . represents a missing value for a variable. The ordering of the variables corresponds to the order of the variables in the analysis. The first row in the display represents the pattern for 26 cases and has X's for all variables but the last (*CALORIES*); for 26 cases, *CALORIES* is the only missing value. Fifty-nine cases have no missing values. *LIT\_FEMA* and *LIT\_MALE* are the only missing values for 15 cases and five cases are missing *CALORIES*, *LIT\_FEMA*, and *LIT\_MALE*. The remaining four cases exhibit unique missing value patterns.



## EM Method

Here we employ the EM algorithm to iteratively arrive at final correlation estimates. This method often performs better than the other methods when data are jointly missing.

The input is:

```
USE WORLD95M
LET LOG DEA = L10(DEATH_RT)
FORMAT 6,3
MISSING
NOTE 'EM Method'
MODEL LOG POP LOG DEN LIFEEXPF LIFEEXPM POP INCR,
BABYMORT LOG GDP BIRTH_RT LOG DEA B_TO_D,
FERTILTY URBAN LITERACY LIT_FEMA,
LIT MALE CALORIES
ESTIMATE 7 MATRIX=CORRELATION ITER=200
```

The output is:

### EM Method

#### EM Algorithm

Iteration	Maximum Error	-2*LL
1	0.982	4.071E+003
2	0.117	4.010E+003
3	0.056	3.983E+003
4	0.028	3.971E+003
5	0.014	3.965E+003
6	0.008	3.962E+003
7	0.005	3.961E+003
8	0.003	3.961E+003
9	0.002	3.961E+003
10	0.002	3.961E+003
11	0.001	3.961E+003
12	0.001	3.961E+003

Mahalanobis D^2 and z-score  
NOTE:

```
Case is an outlier. Mahalanobis D^2 : 38.437 Z : 3.149
Case is an outlier. Mahalanobis D^2 : 67.453 Z : 5.340
Case is an outlier. Mahalanobis D^2 : 37.961 Z : 3.102
Case is an outlier. Mahalanobis D^2 : 69.508 Z : 5.478
Case is an outlier. Mahalanobis D^2 : 38.723 Z : 3.176
Case is an outlier. Mahalanobis D^2 : 39.696 Z : 3.120
```

#### Missing Value Patterns

N of Cases	Missing Value Patterns (X=nonmissing; .=missing)
26	XXXXXXXXXXXXX- X.
59	XXXXXXXXXXXXX- XX
15	XXXXXXXXXXXXX.-



## Missing Value Analysis

```

.X
5 XXXXXXXXXXXX.-
..
1 XXXXXXXXXXXX.XX.-
..
1 XXXXXXXXXXXX...-
.X
1 XXXXXXXXXXXX..-
..
1 XXXXXXXX...XX.-
..

```

Little MCAR Test Statistic : 1.335E+002  
df : 88  
p-value : 0.001

## EM Estimate of Means

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP	BIRTH_RT
	4.114	1.784	70.156	64.917	1.682	42.313	3.422	25.923

## EM Estimate of Means (contd...)

	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA	LIT_MALE	CALORIES
	0.941	3.200	3.530	56.640	78.408	72.717	82.700	2.790E+003

## EM Estimated Correlation Matrix

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT	LOG_GDP
LOG_POP	1.000						
LOG_DEN	0.143	1.000					
LIFEEXPF	-0.088	0.126	1.000				
LIFEEXPM	-0.082	0.153	0.982	1.000			
POP_INCR	-0.078	-0.252	-0.579	-0.502	1.000		
BABYMORT	0.109	-0.152	-0.962	-0.936	0.602	1.000	
LOG_GDP	-0.217	0.004	0.831	0.805	-0.557	-0.824	1.000
BIRTH_RT	-0.027	-0.216	-0.862	-0.805	0.861	0.865	-0.769
LOG_DEA	0.087	-0.070	-0.588	-0.640	-0.202	0.535	-0.323
B_TO_D	-0.153	-0.112	-0.088	-0.012	0.799	0.119	-0.209
FERTILTY	-0.055	-0.233	-0.839	-0.785	0.841	0.835	-0.692
URBAN	-0.138	0.018	0.743	0.729	-0.377	-0.718	0.754
LITERACY	-0.045	0.094	0.864	0.807	-0.700	-0.898	0.727
LIT_FEMA	0.006	0.139	0.838	0.775	-0.703	-0.856	0.686
LIT_MALE	0.070	0.167	0.799	0.748	-0.686	-0.824	0.669
CALORIES	0.009	0.094	0.748	0.732	-0.582	-0.750	0.810

## EM Estimated Correlation Matrix (contd...)

	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY	LIT_FEMA
LOG_POP							
LOG_DEN							
LIFEEXPF							
LIFEEXPM							
POP_INCR							
BABYMORT							
LOG_GDP							
BIRTH_RT	1.000						
LOG_DEA	0.233	1.000					
B_TO_D	0.483	-0.689	1.000				
FERTILTY	0.975	0.275	0.454	1.000			
URBAN	-0.629	-0.427	-0.037	-0.606	1.000		
LITERACY	-0.868	-0.370	-0.280	-0.860	0.646	1.000	
LIT_FEMA	-0.855	-0.325	-0.306	-0.858	0.653	0.970	1.000
LIT_MALE	-0.819	-0.295	-0.310	-0.818	0.630	0.944	0.965
CALORIES	-0.730	-0.213	-0.261	-0.664	0.646	0.637	0.600

## EM Estimated Correlation Matrix (contd...)

	LIT_MALE	CALORIES
LOG_POP		
LOG_DEN		
LIFEEXPF		
LIFEEXPM		
POP_INCR		
BABYMORT		
LOG_GDP		
BIRTH_RT		
LOG_DEA		
B_TO_D		
FERTILTY		
URBAN		
LITERACY		
LIT_FEMA		
LIT_MALE	1.000	
CALORIES	0.635	1.000

## Pairwise Frequency Table

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT
LOG_POP	1.090E+002					
LOG_DEN	1.090E+002	1.090E+002				
LIFEEXPF	1.090E+002	1.090E+002	1.090E+002			
LIFEEXPM	1.090E+002	1.090E+002	1.090E+002	1.090E+002		
POP_INCR	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002	
BABYMORT	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002
LOG_GDP	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002
BIRTH_RT	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002	1.090E+002
LOG_DEA	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002
B_TO_D	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.080E+002
FERTILTY	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002
URBAN	1.080E+002	1.080E+002	1.080E+002	1.080E+002	1.070E+002	1.080E+002
LITERACY	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002
LIT_FEMA	85.000	85.000	85.000	85.000	85.000	85.000
LIT_MALE	85.000	85.000	85.000	85.000	85.000	85.000
CALORIES	75.000	75.000	75.000	75.000	75.000	75.000

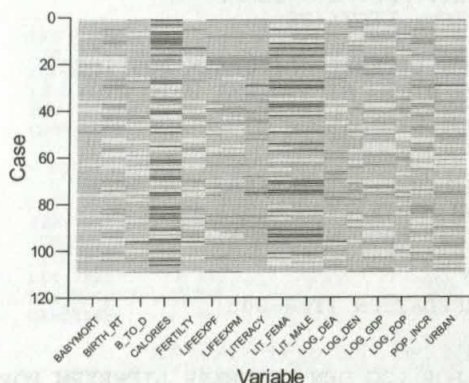
## Pairwise Frequency Table (contd...)

	LOG_GDP	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN
LOG_POP						
LOG_DEN						
LIFEEXPF						
LIFEEXPM						
POP_INCR						
BABYMORT						
LOG_GDP	1.090E+002					
BIRTH_RT	1.090E+002	1.090E+002				
LOG_DEA	1.080E+002	1.080E+002	1.080E+002			
B_TO_D	1.080E+002	1.080E+002	1.080E+002	1.080E+002		
FERTILTY	1.070E+002	1.070E+002	1.070E+002	1.070E+002	1.070E+002	
URBAN	1.080E+002	1.080E+002	1.070E+002	1.070E+002	1.060E+002	1.080E+002
LITERACY	1.070E+002	1.070E+002	1.060E+002	1.060E+002	1.050E+002	1.070E+002
LIT_FEMA	85.000	85.000	85.000	85.000	85.000	85.000
LIT_MALE	85.000	85.000	85.000	85.000	85.000	85.000
CALORIES	75.000	75.000	75.000	75.000	75.000	74.000

Pairwise Frequency Table (contd...)

	LITERACY	LIT_FEMA	LIT_MALE	CALORIES
LOG_POP				
LOG_DEN				
LIFEEXPF				
LIFEEXPM				
POP_INCR				
BABYMORT				
LOG_GDP				
BIRTH_RT				
LOG_DEA				
B_TO_D				
FERTILTY				
URBAN				
LITERACY	1.070E+002			
LIT_FEMA	85.000	85.000		
LIT_MALE	85.000	85.000	85.000	
CALORIES	74.000	59.000	59.000	75.000

Missing Values Plot



Roderick J. A. Little's chi-square statistic for testing whether values are missing completely at random accompanies EM matrices. This statistic has an asymptotic chi-square distribution with degrees of freedom equal to the sum of the number of observed variables across missing value patterns minus the number of variables. In this example, the degrees of freedom equal  $15 + 16 + 14 + 13 + 12 + 12 + 12 + 10 - 16$ , or 88. For a chi-square distribution with 88 degrees of freedom, the obtained value of 133.476 has a *p*-value of .001. This small *p*-value suggests that the missing values are not missing completely at random, but instead depend on the variables in the analysis.

SYSTAT identifies six cases as outliers. Outliers have undue influence on the estimates and you should examine these cases for possible omission from the analysis.

### Example 4

#### Comparing Correlation Estimation Methods

In a large study, it is difficult to compare two correlation matrices for differences (or to determine whether they differ at all). Here, we save three correlation matrices and use MATRIX to compute the differences between elements in each pair of matrices.

The input is:

```

USE WORLD95M
LET LOG_DEA = L10(DEATH_RT)
FORMAT 6,3
CORR
SAVE LCORR
PEARSON LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
        BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
        FERTILTY URBAN LITERACY LIT_FEMA,
        LIT_MALE CALORIES / LISTWISE
SAVE PCORR
PEARSON LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
        BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
        FERTILTY URBAN LITERACY LIT_FEMA,
        LIT_MALE CALORIES / PAIRWISE
MISSING
MODEL LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
        BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
        FERTILTY URBAN LITERACY LIT_FEMA,
        LIT_MALE CALORIES
SAVE EMCORR
ESTIMATE / MATRIX=CORRELATION ITER=200

USE EMCORR/MAT=EMCORR
ROWNAME EMCORR = LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
        BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
        FERTILTY URBAN LITERACY LIT_FEMA,
        LIT_MALE CALORIES
USE PCORR/MAT=PCORR
ROWNAME PCORR = LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
        BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
        FERTILTY URBAN LITERACY LIT_FEMA,
        LIT_MALE CALORIES
USE LCORR/MAT=LCORR
MAT DIFF_LP=LCORR-PCORR
MAT DIFF_LE=LCORR-EMCORR
MAT DIFF_PE=PCORR-EMCORR
SHOW DIFF_LP DIFF_LE DIFF_PE

```



The differences between the listwise and pairwise estimates follow:

Matrix Name: diff\_lp

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT
LOG_POP	0.000	.	.	.	.	.
LOG_DEN	0.139	0.000	.	.	.	.
LIFEEXPF	0.126	-0.122	0.000	.	.	.
LIFEEXPM	0.141	-0.129	0.005	0.000	.	.
POP_INCR	-0.221	0.046	0.188	0.177	0.000	.
BABYMORT	-0.118	0.115	0.011	0.005	-0.182	0.000
LOG_GDP	0.078	-0.220	-0.066	-0.069	0.194	0.079
BIRTH_RT	-0.196	0.080	0.045	0.032	-0.086	-0.057
LOG_DEA	-0.059	0.050	-0.214	-0.183	0.104	0.208
B_TO_D	-0.116	0.039	0.357	0.329	-0.108	-0.349
FERTILTY	-0.180	0.081	0.048	0.035	-0.085	-0.049
URBAN	-0.003	-0.241	-0.003	-0.013	0.183	0.013
LITERACY	0.132	-0.081	-0.039	-0.024	0.132	0.010
LIT_FEMA	0.104	-0.042	-0.004	0.029	0.059	-0.013
LIT_MALE	0.100	-0.042	-0.023	0.010	0.077	0.004
CALORIES	0.097	-0.062	-0.059	-0.054	0.216	0.076

	LOG_GDP	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY
LOG_POP	.	.	.	.	.	.	.
LOG_DEN	.	.	.	.	.	.	.
LIFEEXPF	.	.	.	.	.	.	.
LIFEEXPM	.	.	.	.	.	.	.
POP_INCR	.	.	.	.	.	.	.
BABYMORT	.	.	.	.	.	.	.
LOG_GDP	0.000	.	.	.	.	.	.
BIRTH_RT	0.095	0.000	.	.	.	.	.
LOG_DEA	-0.156	0.238	0.000	.	.	.	.
B_TO_D	0.291	-0.295	-0.041	0.000	.	.	.
FERTILTY	0.108	-0.007	0.235	-0.300	0.000	.	.
URBAN	0.032	0.063	-0.153	0.293	0.086	0.000	.
LITERACY	-0.090	0.047	-0.205	0.314	0.052	-0.035	0.000
LIT_FEMA	-0.030	0.024	-0.128	0.180	0.020	0.022	-0.010
LIT_MALE	-0.031	0.037	-0.116	0.182	0.036	0.008	-0.009
CALORIES	-0.045	0.104	-0.140	0.279	0.115	-0.018	-0.106

	LIT_FEMA	LIT_MALE	CALORIES
LOG_POP	.	.	.
LOG_DEN	.	.	.
LIFEEXPF	.	.	.
LIFEEXPM	.	.	.
POP_INCR	.	.	.
BABYMORT	.	.	.
LOG_GDP	.	.	.
BIRTH_RT	.	.	.
LOG_DEA	.	.	.
B_TO_D	.	.	.
FERTILTY	.	.	.
URBAN	.	.	.
LITERACY	.	.	.
LIT_FEMA	0.000	.	.
LIT_MALE	-0.005	0.000	.
CALORIES	0.000	0.000	0.000

We find many large differences between the correlations estimated by the two deletion methods. The differences are particularly large for  $B\_TO\_D$ .

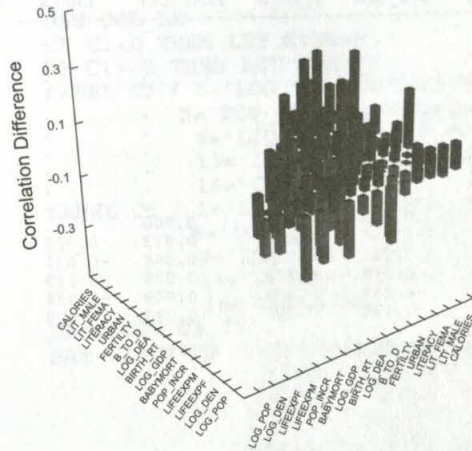
To assist in identifying the large differences, we use MATRIX to create a rectangular data file of correlation differences. We then create a bar chart of these differences.

```

USE PCORR/MATR=PCORR
USE LCORR/MAT=LCORR
MAT DIFF_LP=LCORR-PCORR
CLEAR PCORR LCORR
MAT CIX=[1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16]
MAT CIX=CIX//CIX//CIX//CIX//CIX//CIX//CIX//CIX//
CIX//CIX//CIX //CIX//CIX//CIX//CIX
MAT RIX=TRP(CIX)
MAT RIX=SHAPE(RIX,256,1)
MAT CIX=SHAPE(CIX,256,1)
MAT DIFF_LP=FOLD(DIFF_LP)
MAT COL_LP=SHAPE(DIFF_LP,256,1)
MAT COL_LP=COL_LP||RIX||CIX
SHOW COL_LP
COLNAME COL_LP=C1 C2 C3
MSAVE COL_LP

USE COL_LP
IF C1 < 0 THEN LET SIGN=0
IF C1 >= 0 THEN LET SIGN=1
LABEL C2/1='LOG_POP' 2='LOG_DEN' 3='LIFEEXPF' 4='LIFEEXPM',
5='POP_INCR' 6='BABYMORT' 7='LOG_GDP' 8='BIRTH_RT',
9='LOG_DEA' 10='B_TO_D' 11='FERTILTY' 12='URBAN',
13='LITERACY' 14='LIT_FEMA' 15='LIT_MALE',
16='CALORIES'
LABEL C3/1='LOG_POP' 2='LOG_DEN' 3='LIFEEXPF' 4='LIFEEXPM',
5='POP_INCR' 6='BABYMORT' 7='LOG_GDP' 8='BIRTH_RT',
9='LOG_DEA' 10='B_TO_D' 11='FERTILTY' 12='URBAN',
13='LITERACY' 14='LIT_FEMA' 15='LIT_MALE',
16='CALORIES'
CATEGORY C2 C3
BAR C1*C3*C2 /GROUP=SIGN OVERLAY COLOR=RED,BLUE,
BASE=0 BTHICK= 0.80 LEGEND=NONE,
XLAB='' YLAB='' ZMIN=-.5 ZMAX=.5,
ZLAB='Correlation Difference'

```



The order of the variables along an axis corresponds to variables with little or no missing data at the left end (*LOG\_POP*) and variables with the most missing data at the right end (*CALORIES*). The bar graph reveals that *LOG\_DEA* pairwise correlation estimates tend to be larger than listwise estimates when the variable being correlated with *LOG\_DEA* contains many missing values. The reverse pattern occurs for *B\_TO\_D*. These patterns suggest that the data are not missing completely at random.

### Listwise Deletion vs EM Method

The differences between the listwise and EM correlation estimates follow:

Matrix Name: diff\_le

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT
LOG_POP	0.000					
LOG_DEN	0.139	0.000				
LIFEEXPF	0.126	-0.122	0.000			
LIFEEXPM	0.141	-0.129	0.005	0.000		
POP_INCR	-0.221	0.046	0.188	0.177	0.000	
BABYMORT	-0.118	0.115	0.011	0.005	-0.182	0.000
LOG_GDP	-0.058	-0.220	-0.066	-0.069	0.194	0.079
BIRTH_RT	-0.196	0.080	0.045	0.032	-0.086	-0.057
LOG_DEA	-0.058	0.055	-0.214	-0.183	0.101	0.207
B_TO_D	-0.116	0.040	0.358	0.330	-0.107	-0.350
URBAN	-0.185	0.090	0.049	0.038	-0.086	-0.051
FERTILTY	-0.003	-0.243	-0.002	-0.012	0.185	0.013
LIT_FEMA	0.126	-0.090	-0.038	-0.022	0.133	0.008
LIT_MALE	0.103	-0.068	-0.023	-0.002	0.123	0.000
CALORIES	0.106	-0.070	-0.045	-0.021	0.144	0.019
	0.133	-0.106	-0.032	-0.022	0.189	0.048



Again, we find large differences between many correlations involving *B\_TO\_D*. The EM estimates tend to be larger when values are not missing. *LOG\_DEA* also exhibits large differences, but not to the degree of *B TO D*.

```

USE LCORR/MAT=LCORR
USE EMCORR/MAT=EMCORR
MAT DIFF LE=LCORR-EMCORR
CLEAR LCORR EMCORR
MAT CIX=[1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16]
MAT CIX=CIX//CIX//CIX//CIX//CIX//CIX//CIX//CIX//CIX//
CIX//CIX//CIX//CIX//CIX//CIX
MAT RIX=TRP(CIX)
MAT RIX=SHAPE(RIX,256,1)
MAT CIX=SHAPE(CIX,256,1)
MAT DIFF LE=FOLD(DIFF_LE)
MAT COL_LE=SHAPE(DIFF_LE,256,1)
MAT COL_LE=COL_LE||RIX||CIX
SHOW COL_LE
COLNAME COL_LE=C1 C2 C3
MSAVE COL_LE

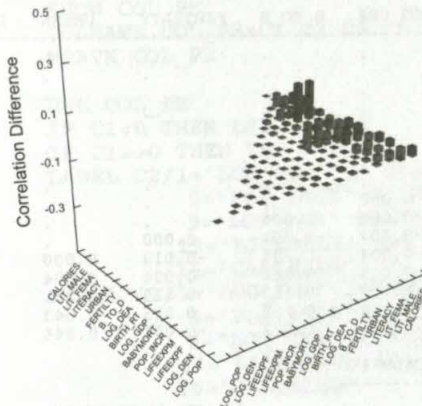
```



```

USE COL LE
IF C1<0 THEN LET SIGN=0
IF C1>=0 THEN LET SIGN=1
LABEL C2 / 1='LOG_POP' 2='LOG_DEN' 3='LIFEEXPF' 4='LIFEEXPM',
           5='POP_INCR' 6='BABYMORT' 7='LOG_GDP' 8='BIRTH_RT',
           9='LOG_DEA' 10='B_TO_D' 11='FERTILTY' 12='URBAN',
           13='LITERACY' 14='LIT_FEMA' 15='LIT_MALE',
           16='CALORIES'
LABEL C3 / 1='LOG_POP' 2='LOG_DEN' 3='LIFEEXPF' 4='LIFEEXPM',
           5='POP_INCR' 6='BABYMORT' 7='LOG_GDP' 8='BIRTH_RT',
           9='LOG_DEA' 10='B_TO_D' 11='FERTILTY' 12='URBAN',
           13='LITERACY' 14='LIT_FEMA' 15='LIT_MALE',
           16='CALORIES'
CATEGORY C2 C3
BAR C1*C3*C2 / GROUP=SIGN OVERLAY COLOR=RED,BLUE,
                BASE=0 BTHICK= 0.80 LEGEND=NONE,
                XLAB='' YLAB='' ZMIN=-.5 ZMAX=.5,
                XLAB='' YLAB='' ZMIN=-.5 ZMAX=.5,
                ZLAB='Correlation Difference'

```



As found elsewhere for pairwise estimates, this bar graph reveals that *LOG\_DEA* EM correlation estimates tend to be larger than listwise estimates when the variable being correlated with *LOG\_DEA* contains many missing values. *B\_TO\_D* exhibits the opposite pattern. For a given pair of variables, the difference between EM and listwise estimates tends to be larger than the difference between pairwise and listwise estimates.

**Pairwise Deletion vs EM Method**

The differences between the pairwise and EM correlation estimates follow:

Matrix Name: diff\_pe

	LOG_POP	LOG_DEN	LIFEEXPF	LIFEEXPM	POP_INCR	BABYMORT
LOG_POP	0.000	.	.	.	.	.
LOG_DEN	0.000	0.000	.	.	.	.
LIFEEXPF	0.000	0.000	0.000	.	.	.
LIFEEXPM	0.000	0.000	0.000	0.000	.	.
POP_INCR	0.000	0.000	0.000	0.000	0.000	.
BABYMORT	0.000	0.000	0.000	0.000	0.000	0.000
LOG_GDP	0.000	0.000	0.000	0.000	0.000	0.000
BIRTH_RT	0.000	0.000	0.000	0.000	0.000	0.000
LOG_DEA	0.001	0.005	0.001	0.000	-0.003	-0.001
B_TO_D	0.000	0.001	0.001	0.001	0.001	-0.001
FERTILTY	-0.005	0.009	0.001	0.002	-0.001	-0.002
URBAN	0.000	-0.002	0.001	0.001	0.002	0.000
LITERACY	-0.005	-0.009	0.001	0.002	0.000	-0.002
LIT_FEMA	-0.001	-0.026	-0.019	-0.031	0.064	0.013
LIT_MALE	0.006	-0.028	-0.022	-0.031	0.067	0.015
CALORIES	0.037	-0.044	0.027	0.033	-0.027	-0.027

	LOG_GDP	BIRTH_RT	LOG_DEA	B_TO_D	FERTILTY	URBAN	LITERACY
LOG_POP	.	.	.	.	.	.	.
LOG_DEN	.	.	.	.	.	.	.
LIFEEXPF	.	.	.	.	.	.	.
LIFEEXPM	.	.	.	.	.	.	.
POP_INCR	.	.	.	.	.	.	.
BABYMORT	.	.	.	.	.	.	.
LOG_GDP	0.000	.	.	.	.	.	.
BIRTH_RT	0.000	0.000	.	.	.	.	.
LOG_DEA	0.001	-0.002	0.000	.	.	.	.
B_TO_D	0.000	0.001	-0.001	0.000	.	.	.
FERTILTY	-0.001	0.000	-0.007	-0.001	0.000	.	.
URBAN	0.001	0.000	-0.004	0.005	-0.013	0.000	.
LITERACY	0.005	-0.001	-0.015	0.009	-0.006	0.004	0.000
LIT_FEMA	-0.053	0.020	-0.117	0.159	0.019	-0.041	0.004
LIT_MALE	-0.058	0.025	-0.118	0.157	0.023	-0.043	0.004
CALORIES	0.038	-0.032	-0.054	0.022	-0.031	0.046	0.044

	LIT_FEMA	LIT_MALE	CALORIES
LOG_POP	.	.	.
LOG_DEN	.	.	.
LIFEEXPF	.	.	.
LIFEEXPM	.	.	.
POP_INCR	.	.	.
BABYMORT	.	.	.
LOG_GDP	.	.	.
BIRTH_RT	.	.	.
LOG_DEA	.	.	.
B_TO_D	.	.	.
FERTILTY	.	.	.
URBAN	.	.	.
LITERACY	.	.	.
LIT_FEMA	0.000	.	.
LIT_MALE	0.000	0.000	.
CALORIES	-0.052	-0.059	0.000

The differences between these two sets of correlation estimates are very small. The largest differences appear for variables missing 22% of the data, *LIT\_FEMA* and *LIT\_MALE*.

As done for the other method comparisons, here we create a bar chart of the correlation differences between the pairwise and EM estimates.

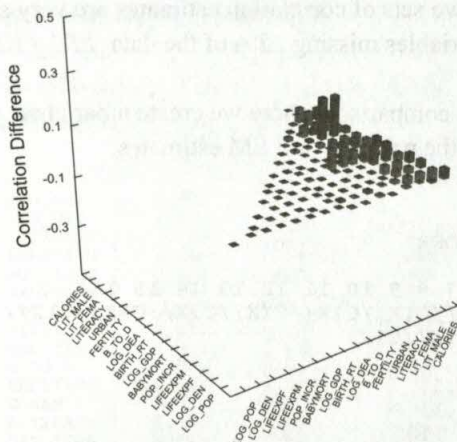
```

USE PCORR/MAT=PCORR
USE EMCORR/MAT=EMCORR
MAT DIFF_PE=PCORR-EMCORR
CLEAR PCORR EMCORR
MAT CIX=[1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16]
MAT CIX=CIX//CIX//CIX//CIX//CIX//CIX//CIX//CIX//CIX//
CIX//CIX//CIX//CIX//CIX//CIX
MAT RIX=TRP(CIX)
MAT RIX=SHAPE(RIX,256,1)
MAT CIX=SHAPE(CIX,256,1)
MAT DIFF_PE=FOLD(DIFF_PE)
MAT COL_PE=SHAPE(DIFF_PE,256,1)
MAT COL_PE=COL_PE||RIX||CIX
SHOW COL_PE
COLNAME COL_PE=C1 C2 C3
MSAVE COL_PE

USE COL_PE
IF C1<0 THEN LET SIGN=0
IF C1>=0 THEN LET SIGN=1
LABEL C2/1='LOG_POP' 2='LOG_DEN' 3='LIFEEXPF' 4='LIFEEXPM',
5='POP_INCR' 6='BABYMORT' 7='LOG_GDP' 8='BIRTH_RT',
9='LOG_DEA' 10='B_TO_D' 11='FERTILTY' 12='URBAN',
13='LITERACY' 14='LIT_FEMA' 15='LIT_MALE',
16='CALORIES'
LABEL C3/1='LOG_POP' 2='LOG_DEN' 3='LIFEEXPF' 4='LIFEEXPM',
5='POP_INCR' 6='BABYMORT' 7='LOG_GDP' 8='BIRTH_RT',
9='LOG_DEA' 10='B_TO_D' 11='FERTILTY' 12='URBAN',
13='LITERACY' 14='LIT_FEMA' 15='LIT_MALE',
16='CALORIES'
CATEGORY C2 C3
BAR C1*C3*C2/GROUP=SIGN OVERLAY COLOR=RED,BLUE,
BASE=0 BTHICK= 0.80 LEGEND=NONE,
XLAB='' YLAB='' ZMIN=-.5 ZMAX=.5,
ZLAB='Correlation Difference'

```





Notice the large empty area in the lower left of the plot. This area corresponds to variables with no missing data; pairwise deletion and EM estimation behave identically in this region. For variables with missing data, the differences between the two estimates are small. The largest differences occur for *LIT\_FEMA* and *LIT\_MALE*.

### Example 5 Missing Value Imputation

MISSING provides EM and regression methods for estimating (imputing) replacement values, but this should not be done until the data have been screened for recording errors and variables in need of a symmetrizing transformation.

Values in the *WORLD95m* data are not randomly missing (we're sure that they are not missing *completely* at random and also have doubts about satisfying the MAR condition). So, how good are the imputed values? In this section, we display some plots that you might create when evaluating your own filled-in data. You can:

- Display the variables with the most values missing in a pair of bivariate scatterplots with the same plot scales—one using the observed data only and the other using the imputed values. For our example, we use *calories* and *lit\_fema*.
- For the same variable, plot the imputed values from one method against those from another. For female literacy, we plot imputed values from the regression method with random residuals against those from the EM method.



**Generating pattern variables.** When evaluating imputation estimates, pattern variables are used as case selection variables to group and identify observed and imputed values. Use the original data to generate pattern variables and merge the pattern variables with the imputed data. Here, we compute pattern variables for calories and female literacy.

```
USE WORLD95M
LET PAT_CAL = CALORIES
LET PAT_LITF = LIT_FEMA
LET (PAT_CAL, PAT_LITF) = @ = .
LET PAT_BOTH = 10*PAT_CAL + PAT_LITF
```

*PAT\_CAL* and *PAT\_LITF* are binary variables. A 1 indicates a missing value and a 0 indicates an observed value. We also generate a third pattern variable (*PAT\_BOTH*) that combines the missing/present information for calories and female literacy. The result of this transformation is four codes: 0, 1, 10, and 11. For example, if, for a case, both values are missing (*PAT\_CAL* and *PAT\_LITF* are both 1), the value of the new variable *PAT\_BOTH* is  $10*1 + 1$  or 11. When only female literacy is missing, the code for *PAT\_BOTH* is 1; when only calories is missing, the code is 10; and when values of both variables are present, the code is 0.

### *Scatterplots of Observed and Imputed Values*

Comparing estimates for variables with many missing values assists in evaluating the performance of the imputation methods. We create pattern variables for *CALORIES* and *LIT\_FEMA* and use them to look for trends in the estimates.

```
USE WORLD95M
LET PAT_CAL = CALORIES
LET PAT_LITF = LIT_FEMA
LET (PAT_CAL, PAT_LITF) = @ = .
LET PAT_BOTH = 10*PAT_CAL + PAT_LITF
LET LOG_DEA = L10(DEATH_RT)
DSAVE WORLD95P
MISSING
MODEL LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
      BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
      FERTILTY URBAN LITERACY LIT_FEMA,
      LIT MALE CALORIES
SAVE REGEST / DATA
ESTIMATE / MATRIX=CORRELATION REGRESSION
MODEL LOG_POP LOG_DEN LIFEEXPF LIFEEXPM POP_INCR,
      BABYMORT LOG_GDP BIRTH_RT LOG_DEA B_TO_D,
      FERTILTY URBAN LITERACY LIT_FEMA,
      LIT MALE CALORIES
SAVE EMEST / DATA
```

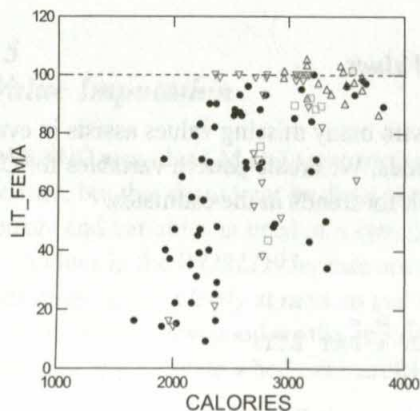
```

ESTIMATE / MATRIX=CORRELATION ITER=200
MERGE WORLD95P (PAT_CAL,PAT_LITF,PAT_BOTH,COUNTRY$) EMEST
DSAVE EMEST2
MERGE WORLD95P (PAT_CAL,PAT_LITF,PAT_BOTH,COUNTRY$) REGEST
DSAVE REGEST2
BEGIN
USE EMEST2
PLOT LIT_FEMA*CALORIES / OVERLAY GROUP=PAT_BOTH YLIMIT=100,
    COLOR=10,2,1,3 SYM=1,4,5,8,
    FILL=1,0,0,0 LEGEND=NONE,
    TITLE='EM Imputed Values',
    LOC=-3in,0in

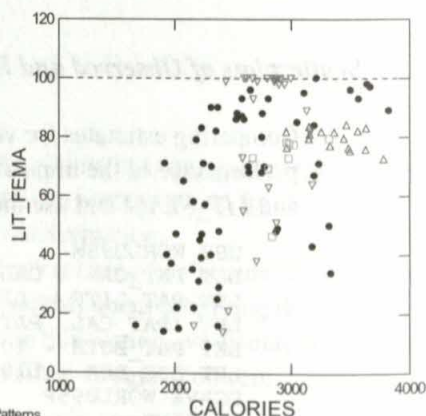
USE REGEST2
PLOT LIT_FEMA*CALORIES / OVERLAY GROUP=PAT_BOTH YLIMIT=100,
    COLOR=10,2,1,3 sym=1,4,5,8,
    LEGEND=-1.6IN,-1.8IN,
    FILL=1,0,0,0 LTITLE='Missing Patterns',
    LLABEL='Both present','LIT missing',
    'CAL missing','Both missing',
    TITLE='Regression Imputed Values',
    LOC=3in,0in
END

```

EM Imputed Values



Regression Imputed Values



Missing Patterns

- Both present
- △ LIT missing
- ▽ CAL missing
- Both missing

Some of the imputed values for both EM and regression lie above 100%. However, the regression estimates tend to be higher. Furthermore, when female literacy is missing, both methods impute values that tend to be high.

### EM vs Regression Imputation

In this example, values imputed by the EM method are compared with those imputed by the regression method. The EM results must be merged with the regression results. To prevent overwriting, we create two new variables in the EM file and merge them with the regression values.

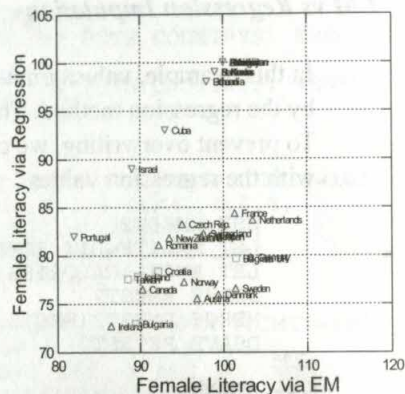
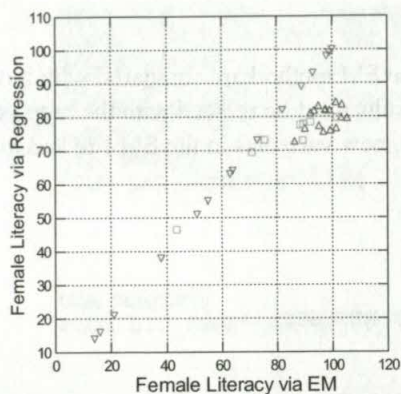
```

USE EMEST2
LET EMLITF=LIT_FEMA
LET EMCAL=CALORIES
DSAVE EMEST2
MERGE EMEST2(EMCAL EMLITF) REGEST2
DSAVE REGEST2

BEGIN
SELECT PAT_BOTH>0
PLOT LIT_FEMA*EMLITF / OVERLAY GROUP=PAT_BOTH COLOR=2,1,3,
                        SYM=4,5,8 FILL=0,0,0 XGRID YGRID,
                        LEGEND=NONE LOC=-3IN,0IN XMAX=120,
                        XLAB='Female Literacy via EM',
                        YLAB='Female Literacy via Regression'
SELECT EMLITF>80 AND PAT_BOTH>0
PLOT LIT_FEMA*EMLITF / OVERLAY GROUP=PAT_BOTH COLOR=2,1,3,
                        SYM=4,5,8 FILL=0,0,0 XGRID YGRID,
                        LEGEND=-1.6IN,-1.8IN,
                        LTITLE='Missing Patterns',
                        LLABEL='LIT missing','CAL missing',
                        'Both missing' LABEL=COUNTRY$,
                        LOC=3IN,0IN XMAX=120,
                        XLAB='Female Literacy via EM',
                        YLAB='Female Literacy via Regression'

END
SELECT PAT_BOTH>0
PLOT CALORIES*EMCAL / OVERLAY GROUP=PAT_BOTH COLOR=2,1,3,
                       SYM=4,5,8 FILL=0,0,0 XGRID YGRID,
                       LTITLE='Missing Patterns',
                       LLABEL='LIT missing','CAL missing',
                       'Both missing' XMAX=4000 YMIN=1500,
                       YMAX=4000 XLAB='Calories via EM',
                       YLAB='Calories via Regression'

```



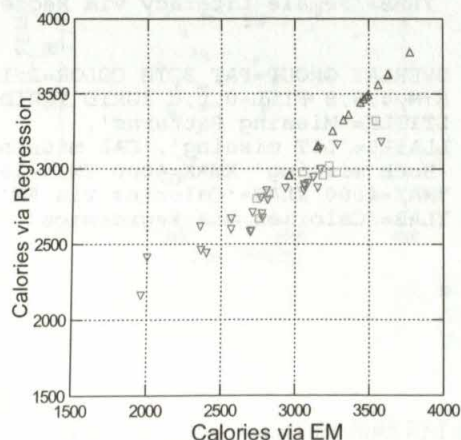
Missing Patterns

△ LIT missing

▽ CAL missing

□ Both missing

Ideally, the points should fall along a line connecting the intersection of grid lines for the same percentage (for example, 80% for EM with 80% for regression). When both calories and female literacy are estimated, the regression estimates tend to be higher than the EM estimates. The points with estimated literacy values are clustered together, making it difficult to identify them in the left plot. On the right side, we zoom in on the area containing the imputed *LIT\_FEMA* values.



Missing Patterns

△ LIT missing

▽ CAL missing

□ Both missing



In this plot, we compare imputed values for *CALORIES*. In general, when there is a difference, the regression estimates tend to be higher more often than they are lower.

### Example 6

#### Regression Imputation

Here, we use a subset of the *WORLD95m* data to illustrate the mechanics underlying regression imputation. Two of the three variables used (*CALORIES* and *LIT\_FEMA*) contain missing values. The third variable, *LOG\_GDP*, is complete. We also create pattern variables for subsequent plotting.

The input is:

```
USE WORLD95M
LET PAT_CAL = CALORIES
LET PAT_LITF = LIT_FEMA
LET (PAT_CAL, PAT_LITF) = @ = .
LET PAT_BOTH = 10*PAT_CAL + PAT_LITF
ESAVE WORLD95M

MISSING
SAVE RESULTS / DATA
MODEL LOG_GDP LIT_FEMA CALORIES
ESTIMATE / MATRIX=CORRELATION REGRESSION
```

The output is:

```
No. of Missing value patterns
Cases (X=nonmissing; .=missing)
  26 XX.
  59 XXX
  16 X.X
   8 X..

Regression Substitution estimate of means
      LOG_GDP      LIT_FEMA      CALORIES
      3.422      70.140      2781.935

Regression Substitution estimated correlation matrix
      LOG_GDP      LIT_FEMA      CALORIES
LOG_GDP      1.000
LIT_FEMA      0.592      1.000
CALORIES      0.810      0.505      1.000
```

Fifty-nine cases contain complete data. Twenty-six cases lack a value for *CALORIES* only, and sixteen cases lack only a *LIT\_FEMA* value. Eight cases are missing data for both *CALORIES* and *LIT\_FEMA*.

### Regression Surfaces

The three patterns involving missing data for at least one variable result in three regression equations for imputing values for the missing entries. Two of these models correspond to simple linear regression:

$$\text{CALORIES} = \beta_0 + \beta_1(\text{LOG\_GDP}) + \beta_2(\text{LIT\_FEMA})$$

$$\text{LIT\_FEMA} = \beta_0 + \beta_1(\text{LOG\_GDP}) + \beta_2(\text{CALORIES})$$

The third model involves a multivariate regression of *CALORIES* and *LIT\_FEMA* on *LOG\_GDP*.

To derive the imputed values, SYSTAT begins by substituting the mean of all available data for each variable for each missing entry. The mean-substituted data yield estimates of the regression coefficients, which can then be used to predict the missing values for each case.

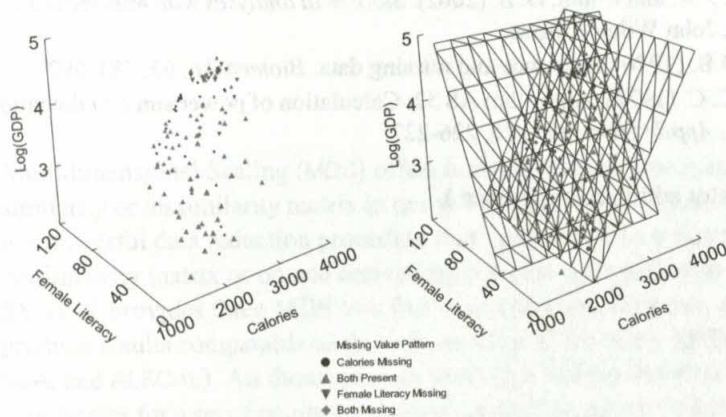
The regression surfaces illustrate the regression imputation procedure. We create side-by-side plots of the imputed data and the regression surfaces:

```
MERGE WORLD95M.SYD (PAT_CAL,PAT_LITF,PAT_BOTH) RESULTS.SYD
LABEL PAT_BOTH / 0='Both Present',
                  1='Female Literacy Missing',
                  10='Calories Missing',
                  11='Both Missing'
ORDER PAT_BOTH / SORT='Both Present',
                  'Female Literacy Missing',
                  'Calories Missing',
                  'Both Missing'
CATEGORY PAT_BOTH
BEGIN
PLOT LOG_GDP*LIT_FEMA*CALORIES / GROUP=PAT_BOTH,
     OVERLAY COLOR=10,1,2,12 SYMBOL=1,4,5,9,
     SIZE=.1,1,1,1, XLABEL="Calories",
     YLABEL='Female Literacy' ZLABEL='Log(GDP)',
     LEGEND=4.3,-2.5 LTITLE = "Missing Value Pattern",
     FILL=1 XMIN=1000,XMAX=4000,YMIN=0,YMAX=120,
     ZMIN=2 ZMAX=5 LOC=-3.3IN,0IN
PLOT LOG_GDP*LIT_FEMA*CALORIES / GROUP=PAT_BOTH,
     OVERLAY COLOR=10,1,2,12 SYMBOL=1,4,5,9,
     SIZE=.1,1,1,1, XLABEL="Calories",
     YLABEL='Female Literacy' ZLABEL='Log(GDP)',
     LEGEND=NONE FILL=1 XMIN=1000 XMAX=4000,
     YMIN=0 YMAX=120 ZMIN=2 ZMAX=5 LOC=3.3IN,0IN
SELECT PAT_BOTH=1
PLOT LOG_GDP*LIT_FEMA*CALORIES / SMOOTH=LINEAR,
     SURFACE=XYCUT COLOR=1 FILL=1 XLABEL="Calories",
     YLABEL='Female Literacy' ZLABEL='Log(GDP)',
```

```

LEGEND=NONE XMIN=1000 XMAX=4000,
YMIN=0 YMAX=120 ZMIN=2 ZMAX=5 LOC=3.3IN,0IN
SELECT PAT BOTH=10
PLOT LOG GDP*LIT FEMA*CALORIES / SMOOTH=LINEAR,
SURFACE=XYCUT COLOR=2 FILL=1 XLABEL="Calories",
YLABEL='Female Literacy' ZLABEL='Log(GDP)',
LEGEND=NONE XMIN=1000 XMAX=4000,
YMIN=0 YMAX=120 ZMIN=2 ZMAX=5 LOC=3.3IN,0IN
END
SELECT

```



Rotating the graphs allows us to view the three-dimensional space from multiple perspectives. When viewing the space along the regression surfaces, we find that many points lie exactly within each regression plane. One plane contains cases with imputed estimates for *CALORIES*, and the other plane contains cases with imputed estimates for *LIT\_FEMA*. These are the planes used to predict the imputed values. Notice that cases lacking values for both *CALORIES* and *LIT\_FEMA*, plotted with a diamond, lie in both regression planes.

## Computation

### Algorithms

The computational algorithms use provisional means, sums of squares, and cross-products (Spicer, 1972). Starting values for the EM algorithm use all available values (see Little and Rubin, 2002).

## References

- Buck, S.F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *Journal of the Royal Statistical Society*, B22, 302-306.
- Little, R.J.A. (1988). Robust estimation of the mean and covariance matrix from data with missing values. *Applied Statistics*, 37, 23-28.
- Little, R. J. A. and Rubin, D. B. (2002). *Statistical analyses with missing data*, 2nd ed. New York: John Wiley & Sons.
- \* Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.
- Spicer, C. C. (1972). Algorithm AS 52: Calculation of power sums of deviations about the mean. *Applied Statistics*, 21, 226-227.

(\* indicates additional reference.)



# Multidimensional Scaling

Leland Wilkinson

Multidimensional Scaling (MDS) offers nonmetric multidimensional scaling of a similarity or dissimilarity matrix in one to five dimensions. Multidimensional scaling is a powerful data reduction procedure that can be used on a direct similarity or dissimilarity matrix or on one derived from rectangular data with Correlations. SYSTAT provides three MDS loss functions (Kruskal, Guttman, and Young) that produce results comparable to those from three of the major MDS packages (KYST, SSA, and ALSCAL). All three methods perform a similar function: to compute coordinates for a set of points in a space such that the distances between pairs of these points fit as closely as possible to measured dissimilarities between a corresponding set of objects.

The family of procedures called *principal components* or *factor analysis* is related to multidimensional scaling in function, but multidimensional scaling differs from this family in important respects. Usually, but not necessarily, multidimensional scaling can fit an appropriate model in fewer dimensions than can these other procedures. Furthermore, if it is implausible to assume a linear relationship between distances and dissimilarities, multidimensional scaling nevertheless provides a simple dimensional model.

MDS also computes the INDSCAL (individual differences multidimensional scaling) model (Carroll and Chang, 1970). The INDSCAL model fits dissimilarity or similarity matrices for multiple subjects into one common space, with jointly estimated weight parameters for each subject (that is, a dissimilarity matrix is input for each subject and separate (monotonic) regression functions are computed). MDS can fit the INDSCAL model using any of the three loss functions, although we recommend using Kruskal's STRESS for this purpose.

Finally, MDS can fit the nonmetric unfolding model. This allows one to analyze rank-order preference data.

## Statistical Background

Multidimensional scaling (MDS) is a procedure for fitting a set of points in a space such that the distances between points correspond as closely as possible to a given set of dissimilarities between a set of objects. Dissimilarities may be measured directly, as in psychological judgments, or derived indirectly as in correlation matrices computed on rectangular data.

### Assumptions

Because MDS, like cluster analysis, operates directly on dissimilarities, no statistical distribution assumptions are necessary. There are, however, other important assumptions. First, multidimensional scaling is a spatial model. To fit points in the kind of spaces that MDS covers, assume that your data satisfy *metric* conditions:

- The distance from an object to itself is 0,
- The distance from object A to object B is the same as that from B to A,
- The distance from object A to C is less than or equal to the distance from A to B plus B to C. This is sometimes called the **triangle inequality**.

You may think these conditions are obvious, but there are numerous counter-examples in psychological perception and elsewhere. For example, commuters often view the distance from home to the city as closer than the distance from the city to home because of traffic patterns, terrain, and psychological expectations related to time of day. Framing or context effects can also disrupt the metric axioms, as Amos Tversky has shown. For example, Miami is similar to Havana. Havana is similar to Moscow. Is Miami similar to Moscow? If your data (objects) are not consistent with these three axioms, do not use MDS.

Second, there are ways of deriving distances from rectangular data that do not satisfy the metric axioms. The ones available in Correlations do, but if you are thinking of using some other derived measure of similarity, check it carefully.

Finally, it is assumed that all your objects will fit in the same metric space. It is best if they diffuse somewhat evenly through this space as well. Do not expect to get

interpretable results for 25 nearly indistinguishable objects and one that is radically different.

## ***Collecting Dissimilarity Data***

You can collect dissimilarities directly or compute them indirectly.

### ***Direct Methods***

Examples of direct dissimilarities are:

**Distances.** Take distances between objects (for example, cities) directly off a map. If the scale is local, MDS will reproduce the map nicely. If the scale is global, you will need three dimensions for an MDS fit. Two or three dimensional spatial distances can be measured directly. Direct measures of social distance might include spatial propinquity or the number of times or amount of time one individual interacts with another.

**Judgments.** Ask subjects to give a numerical rating of the dissimilarity (for example, 0 to 10) between all pairs of objects.

**Clusters.** Ask people to sort objects into piles; or examine naturally occurring aggregates, such as paragraphs, communities, and associations. Record 0 if two objects occur in the same group and 1 if they do not. Sum these counts over replications or judges.

**Triads.** Ask subjects to compare three objects at a time and report which two are most similar (or which is the odd one out). Do this over all possible triads of objects. To compute dissimilarities, sum over all triads, as for the clustering method. There are usually many more triads than pairs of objects, so this method is more tedious. However, it allows you to independently assess possible violations of the triangle inequality.

### ***Indirect Methods***

Indirect dissimilarities are computed over a rectangular matrix whose columns are objects and rows are attributes. You can transpose this matrix if you want to scale rows instead. Possible indirect dissimilarities include:



**Computed Euclidean distances.** These are the square root of the sum-of-squared discrepancies between columns of the rectangular matrix.

**Negatives of correlations.** For standardized data (mean of 0 and standard deviation of 1), Pearson correlations are proportional to Euclidean distances. For unstandardized data, Pearson correlations are comparable to computing Euclidean distances after standardizing. MDS automatically negates correlations if you do not. Other types of correlations for example, Spearman and gamma are analogous to standardized distances, but only approximately. Also, be aware that large negative correlations will be treated as large distances and large positive correlations as small distances. Make sure that all variables are scored in the same direction before computing correlations. If you find that a whole row of a correlation matrix is negative, reverse the variable by multiplying by  $-1$ , and recompute the correlations.

**Counts of discrepancies.** Counting discrepancies between columns or using some of the binary association measures in Correlations is closely related to computing the Euclidean distance. These methods are also related to the clustering distance calculations mentioned above for direct distances.

## ***Scaling Dissimilarities***

Once you have dissimilarities (or similarities, correlations, etc., which MDS automatically transforms to dissimilarities), you may scale them. You do not need to know how the computer does the calculations in order to use the program intelligently as long as you pay attention to the following:

### ***Stress and Iterations***

**Stress** is the goodness-of-fit statistic that MDS tries to minimize. It consists of the square root of the normalized squared discrepancies between interpoint distances in the MDS plot and the smoothed distances predicted from the dissimilarities. Stress varies between 0 and 1, with values near 0 indicating better fit. It is printed for each *iteration*, which is one movement of all the points in the plot toward a better solution. Make sure that iterations proceed smoothly to a minimum. This is true for the examples in this chapter. If you find that the stress values increase or decrease in uneven steps, you should be suspicious.



### ***The Shepard Diagram***

The Shepard diagram is a scatterplot of the distances between points in the MDS plot against the observed dissimilarities (or similarities). The points in the plot should adhere cleanly to a curve or straight line (which would be the smoothed distances). In other words, you should look at a good Shepard plot and think it resembles the outcome of a well-designed experiment. For more information refer examples in the chapter.

If the Shepard diagram resembles a stepwise or *L*-shaped function, beware, you may have achieved a degenerate solution. Publish it and you will be excoriated by the clergy.

### ***The MDS Plot***

The plot of points is what you seek. The points should be scattered fairly evenly through the space. The orientation of axes is arbitrary—remember we are scaling distances, not axes. Feel free to reverse axes or rotate the solution. MDS rotates it to the largest dimensions of variation, but these do not necessarily mean anything for your data.

You may interpret the axes as in principal components or factor analysis. More often, however, you should look for clusters of objects or regular patterns among the objects, such as circles, curved manifolds, and other structures. See the Guttman loss function example for a good view of a circle.

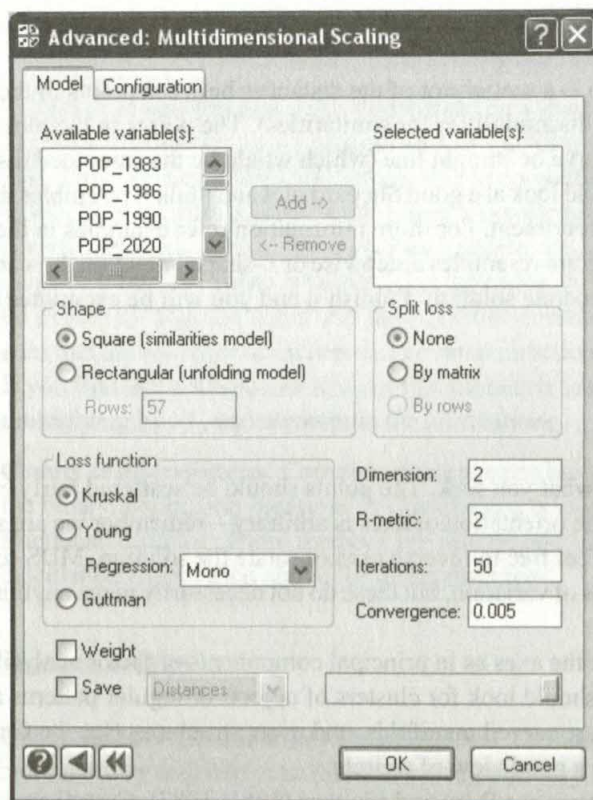
For more information, see Borg and Lingoes (1981, 1987), Carroll and Arabie (1980), Davison (1983), Green and Rao (1972), Kruskal, Wish and Uslander (2006), Schiffman, Reynolds, and Young (1981), and Shepard, Romney and Nerlove (1972).

## ***Multidimensional Scaling in SYSTAT***

### ***Multidimensional Scaling Dialog Box***

To open the Multidimensional Scaling dialog box, from the menus choose:

Advanced  
Multidimensional Scaling...



The following options are available:

**Selected variable(s).** Select the variables that contain the matrix of data to be analyzed.

**Shape.** Specify the type of matrix input. For a similarities model, select Square. For an unfolding model, select Rectangular and enter the number of rows in your matrix.

**Loss function.** MDS scales similarity and dissimilarity matrices using three loss functions:

- **Kruskal.** Uses Kruskal's STRESS formula 1 scaling method.
- **Young.** Uses Young's S-STRESS scaling method, which allows you to scale using the loss function featured in ALSCAL.
- **Guttman.** Uses Guttman's coefficient of alienation scaling method.

**Note:** Iterations with Kruskal's method are faster but usually take longer to converge to a minimum value than those with the Guttman method. The procedure used in the

latter has been found in simulations to be less susceptible to local minima than that used in the Kruskal method (Lingoes and Roskam, 1973). We do not recommend Young's S-STRESS loss function. Because it weights squares of distances, large distances have more influence than smaller ones. Weinberg and Menil (1993) summarized why this is a problem: "...error variances of dissimilarities tend to be positively correlated with their means. If this is the case, large distances should be, if anything, *down-weighted* relative to small distances."

When using the Kruskal or Young loss functions, choose the form of the function relating distances to similarities (or dissimilarities):

- **Mono.** Specifies nonmetric scaling.
- **Linear.** Specifies metric scaling.
- **Log.** Specifies a log function, allowing a smooth curvilinear relation between dissimilarities and distances.
- **Power.** Specifies a power function. (This option is available only with Kruskal loss function.)

By default, SYSTAT takes it as Kruskal MONOTONIC loss function.

**Note:** If you use the Kruskal loss function, you can fit a MONOTONIC, LINEAR, or LOG function of distances onto input dissimilarities. The standard option is MONOTONIC multidimensional scaling. To avoid degenerate solutions, however, log or linear scaling is sometimes handy. Log scaling is recommended for this purpose because it allows a smooth curvilinear relation between dissimilarities and distances.

**Split loss.** For an individual differences of unfolding model, split the calculation of the loss function by rows of the matrix or by matrices. Splitting by rows is possible only for a rectangular matrix.

**Dimension.** Number of dimensions in which to scale. The number of dimensions must be a positive integer less than or equal to the number of variables that you scale and 5. The default value is 2.

**R-metric.** Constant for the Minkowski power metric for computing distances. For ordinary Euclidean distance, enter 2. For city-block distance, enter 1. For values other than 1 or 2, computation is slower because logarithms and exponentials are used. The default value is 2.

The general formula for calculating distances is:



$$d_{jk} = \left[ \sum_{i=1}^p |x_{ij} - x_{ik}|^r \right]^{\frac{1}{r}}$$

where  $r$  is the specified power and  $p$  is the number of dimensions.

**Iterations.** Limit for the number of iterations.

**Convergence.** Iterations terminate when the maximum absolute difference between any coordinate in the solution at iteration  $i$  versus iteration  $i - 1$  is less than the specified convergence criterion. Because the configuration is standardized to unit variance on every iteration, iteration stops when no coordinate moves more than the specified convergence criterion (0.005 by default) from its value on the previous iteration.

Most MDS programs terminate when stress reaches a predetermined value or changes by less than a small amount. These programs can terminate prematurely, however, because comparable stress values can result from different configurations. The SYSTAT convergence criterion allows you to stop iterating when the configuration ceases to change.

**Weight.** Adds weights for each dimension and each matrix (subject) into the calculation of separate distances that are used in the minimization. For an individual differences model, select Weight.

**Save.** You can save three sets of output to a data file:

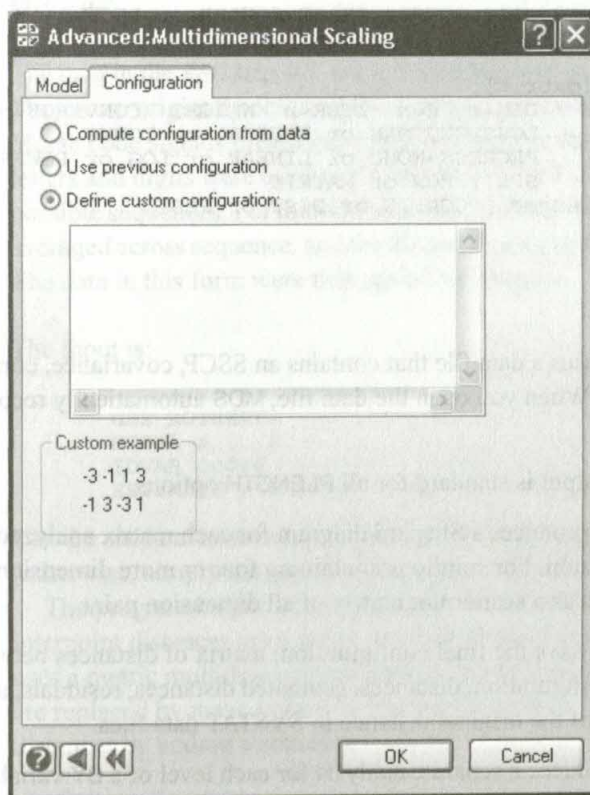
- **Configuration.** Saves the final configuration.
- **Distances.** Saves the matrix of distances between points in the final scaled configuration.
- **Residuals.** Saves the data, distances, estimated distances, residuals, and the row and column number of the original distance in the rectangular SYSTAT file.

With the residuals, MDS displays the root-mean-squared residuals for each point in its output. Because STRESS is a function of the sum-of-squared residuals, the root-mean-squared residuals are a measure of the influence of each point on the STRESS statistic. This can help you identify ill-fitting points.



## Multidimensional Scaling Configuration

SYSTAT offers several alternative initial configurations.



**Compute configuration from data.** By default, the configuration is computed from the data. The method used depends on the loss function.

**Use previous configuration.** Uses the configuration from the previous scaling.

**Define custom configuration.** You can specify a custom starting configuration for the scaling. There must be as many rows as items and columns as dimensions. When you type a matrix, SYSTAT reads as many numbers in each row as you specify. It reads as many rows as there are points to scale.

You can specify a configuration for confirmatory analysis. Enter a hypothesized configuration and let the program iterate only once. Then look at the stress.

## Using Commands

First, specify your data with *USE filename*. Continue with:

```
MDS
  MODEL varlist / ROWS=n SHAPE=SQUARE or RECT
  CONFIG LAST
  or
  CONFIG [matrix]
  ESTIMATE / DIM=n R=n ITER=n WEIGHT CONVERGE=n ,
            LOSS=GUTTMAN or KRUSKAL or YOUNG ,
            REGRESS=MONO or LINEAR or LOG or POWER ,
            SPLIT=ROW or MATRIX
  SAVE filename / CONFIG or DIST or RESID
```

## Usage Considerations

**Types of data.** MDS uses a data file that contains an SSCP, covariance, correlation, or dissimilarity matrix. When you open the data file, MDS automatically recognizes its type.

**Print options.** The output is standard for all PLENGTH options.

**Quick Graphs.** MDS produces a Shepard diagram for each matrix analyzed and a plot of the final configuration. For solutions containing four or more dimensions, the final configuration appears as a scatterplot matrix of all dimension pairs.

**Saving files.** You can save the final configuration, matrix of distances between points in the final scaled configuration, distances, estimated distances, residuals, and the row and column number of the original distance in SYSTAT data files.

**BY groups.** MDS produces a separate analysis for each level of a BY variable.

**Case frequencies.** FREQ is not available in MDS.

**Case weights.** WEIGHT is not available in MDS.

## Examples

### Example 1 Kruskal Method

The data in the *ROTHKOPF* file are adapted from an experiment by Rothkopf (1957). They were originally obtained from 598 subjects who judged whether or not pairs of Morse code signals presented in succession were the same. Morse code signals for letters and digits were used in the experiment, and all pairs were tested in each of two possible sequences. For multidimensional scaling, the data for letter signals have been averaged across sequence, and the diagonal (pairs of the same signal) has been omitted. The data in this form were first scaled by Shepard.

The input is:

```
MDS
  USE ROTHKPF1
  MODEL a .. z
  IDVAR code$
  ESTIMATE / LOSS=KRUSKAL
```

Use the shortcut notation (..) in MODEL for listing consecutive variables in the file (otherwise, simply list each variable name separated by a space).

The program begins by generating an initial configuration of points whose interpoint distances are a linear function of the input data. For this estimation, MDS uses a metric multidimensional scaling. To do this, missing values in the input matrix are replaced by mean values for the whole matrix. Then the values are converted to distances by adding a constant.

The output is:

Monotonic Multidimensional Scaling

Kruskal Method

The data are analyzed as similarities

Minimizing Kruskal STRESS (form 1) in 2 dimensions

Iteration History

Iteration	STRESS
0	0.263539
1	0.237909
2	0.218820
3	0.202184
4	0.190513
5	0.184341
6	0.181174
7	0.179394
8	0.178269

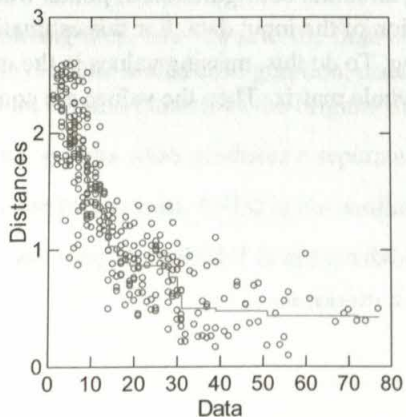
## Chapter 4

Stress of Final Configuration : 0.178269  
 Proportion of Variance (RSQ) : 0.845020

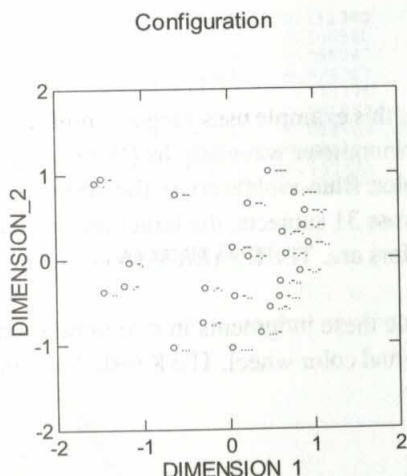
## Coordinates in 2 Dimensions

Variable	Dimension	
	1	2
..	-1.211291	-0.310037
....	0.587818	-0.449746
---	0.667949	0.050103
---	0.061532	-0.439883
---	-1.542846	0.893490
---	0.475856	-0.571910
---	0.224256	0.645882
---	0.032423	-1.047075
---	-1.447269	-0.381961
---	0.776074	0.765947
---	0.224747	0.024567
---	0.603292	-0.269646
---	-0.621882	0.757884
---	-1.153966	-0.042454
---	0.468887	1.024640
---	0.629749	0.305905
---	0.897228	0.555671
---	-0.283513	-0.343725
---	-0.655589	-1.038669
---	-1.469059	0.948010
---	-0.310876	-0.750825
---	0.365593	-0.869607
---	0.041743	0.131315
---	0.832711	-0.148606
---	0.870719	0.381966
---	0.935717	0.178765

Shepard Diagram







The solution required eight iterations. Notice that STRESS reduces at each iteration. Final STRESS values near zero may indicate the presence of a degenerate solution.

The Shepard diagram is a scatterplot of distances between points in the MDS plot against the observed dissimilarities or similarities. In monotonic scaling, the regression function has steps at various points. For most solutions, the function in this plot should be relatively smooth (without large steps). If the function looks like one or two large steps, you should consider setting REGRESSION to LOG or LINEAR under ESTIMATE.

Notice that large values of the data tend to have small distances in the configuration. The diagram displays an overall decreasing trend because we are using similarities (large data values indicate similar objects). For dissimilarities, the Shepard diagram displays an increasing trend.

In the configuration plot, the points should be scattered fairly evenly through the space. If you are scaling in more than two dimensions, you should examine plots of pairs of axes or rotate the solution in three dimensions. The solution has been rotated to principal axes (that is, the major variation is on the first dimension). This rotation is not performed unless the scaling is in Euclidean space, as in the present example.

The two-dimensional solution clearly distinguishes short signals from long and dots from dashes. Dashes tend to appear in the upper right and dots in the lower left. Long codes tend to appear in the lower right and short in the upper left.

## Example 2

### Guttman Loss Function

To illustrate the Guttman loss function, this example uses judged similarities among 14 spectral colors (from Ekman, 1954). Nanometer wavelengths (W434, ..., W674) are used to name the variables for each color. Blue-violets are in the 400's; reds are in the 600's. The judgments are averaged across 31 subjects; the larger the number for a pair of colors, the more similar the two colors are. The file (*EKMAN*) has no diagonal elements, and its type is *SIMILARITY*.

The Guttman method is used to scale these judgments in two dimensions to determine whether the data fit a perceptual color wheel. The Kruskal loss function will give you a similar result.

The input is:

```
MDS
USE EKMAN
MODEL w434 .. w674
ESTIMATE / LOSS=GUTTMAN
```

The output is:

Monotonic Multidimensional Scaling

Guttman loss function  
The data are analyzed as similarities  
Minimizing Guttman/Lingoes Coefficient of Alienation in 2 dimensions

Iteration History

Iteration	Alienation
0	0.070826
1	0.042072
2	0.037764
3	0.036151
4	0.035074

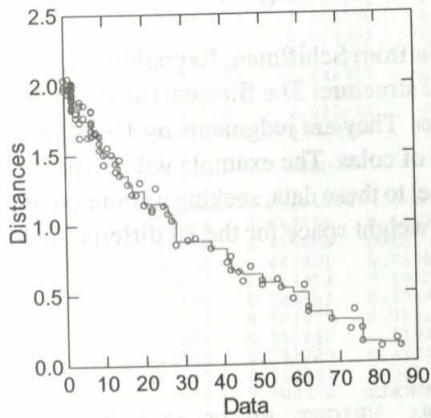
Alienation of Final Configuration : 0.035074  
Proportion of Variance (RSQ) : 0.996227

Coordinates in 2 Dimensions

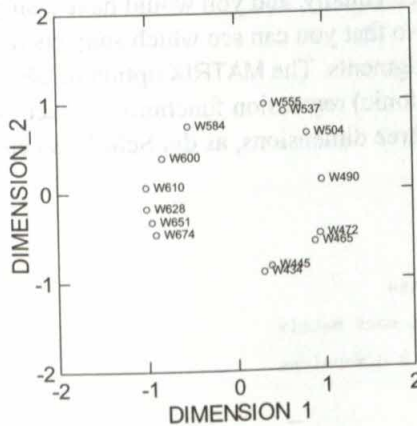
Variable	Dimension	
	1	2
W434	0.311713	-0.905203
W445	0.400413	-0.840312
W465	0.893585	-0.574320
W472	0.952088	-0.484501

W490	0.975491	0.112340
W504	0.814841	0.640540
W537	0.547614	0.888347
W555	0.329882	0.974307
W584	-0.536487	0.734375
W600	-0.826975	0.381875
W610	-1.010004	0.056985
W628	-1.005072	-0.181708
W651	-0.944729	-0.332423
W674	-0.902358	-0.470305

Shepard Diagram



Configuration



The fit of configuration distances to original data is extremely close, as evidenced by the low coefficient of alienation and clean Shepard diagram.

The resulting configuration is almost circular, denoting a "circumplex" by Guttman (1954). There is a large gap at the bottom of the figure, however, because the perceptual color between deep red and dark purple is not a spectral color.

### **Example 3** **Individual Differences Multidimensional Scaling**

The data in the *COLAS* file are taken from Schiffman, Reynolds, and Young (1981). The data in this file have an unusual structure. The file consists of 10 dissimilarity matrices stacked on top of each other. They are judgments by 10 subjects of the dissimilarity (0–100) between pairs of colas. The example will fit the INDSCAL (individual differences scaling) model to these data, seeking a common group space for the 10 different colas and a parallel weight space for the 10 different judges.

The input is:

```
MDS
  USE COLAS
  MODEL dietpeps .. dietrite
  ESTIMATE / LOSS=KRUSKAL WEIGHT SPLIT=MATRIX DIM=3
```

The **WEIGHT** option tells SYSTAT to weight each matrix separately. Without this option, all matrices would be weighted equally, and you would have a single pooled solution. You want to use weighting so that you can see which subjects favor one dimension over the others in their judgments. The **MATRIX** option of **SPLIT** tells SYSTAT to compute separate (monotonic) regression functions for each subject (matrix). Finally, scale the result in three dimensions, as did Schiffman et al. (1981).

The output is:

```
Monotonic Multidimensional Scaling
Kruskal Method
The data are analyzed as dissimilarities
There are 10 replicated data matrices
Dimensions are weighted separately for each matrix
Fitting is split between data matrices
Minimizing Kruskal STRESS (form 1) in 3 dimensions
```

Iteration History



Iteration	STRESS
-----------	--------

0	0.220898
1	0.184422
0	0.221309
1	0.184508

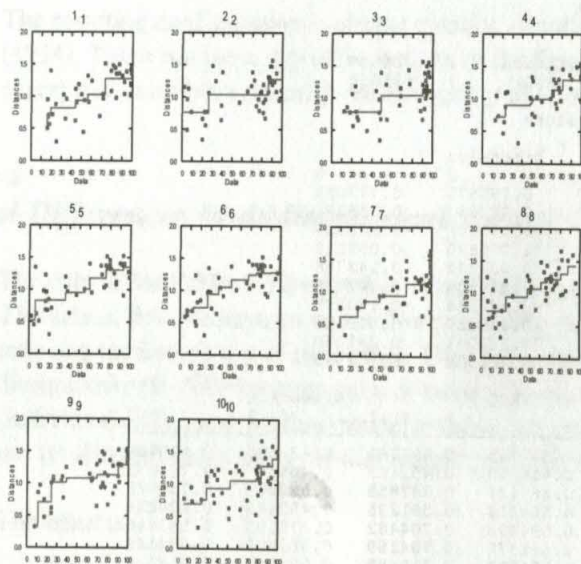
Stress of Final Configuration : 0.184508  
 Proportion of Variance (RSQ) : 0.535014

## Coordinates in 3 Dimensions

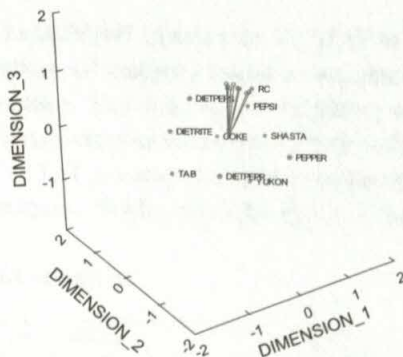
Variable	1	2	3
DIETPEPS	-0.608199	0.195575	0.777055
RC	0.521748	0.052353	0.756390
YUKON	0.415860	-0.089042	-0.867859
PEPPER	0.271872	-1.265870	0.059119
SHASTA	0.797845	0.024902	-0.143788
COKE	0.390732	0.836586	-0.347338
DIETPEPR	-0.747107	-0.842914	-0.173399
TAB	-0.790969	0.438430	-0.609165
PEPSI	0.570666	0.221001	0.381030
DIETRITR	-0.822448	0.428980	0.167955

Matrix	Stress	RSQ	Dimension		
			1	2	3
1	0.188374	0.547755	0.697761	0.433686	0.526788
2	0.199808	0.416268	0.452105	0.465357	0.721233
3	0.196430	0.467821	0.347855	0.522989	0.739323
4	0.170677	0.564314	0.591235	0.492475	0.608234
5	0.178156	0.594393	0.704482	0.370109	0.563840
6	0.171913	0.621371	0.704169	0.367609	0.570188
7	0.181071	0.551692	0.419485	0.582263	0.659122
8	0.180465	0.559729	0.483517	0.597254	0.608641
9	0.163263	0.624525	0.562688	0.495564	0.625805
10	0.211658	0.402270	0.435248	0.609372	0.617438

## Shepard Diagram



Configuration



The solution required four iterations. Notice that the second two iterations appear to be a restart. This is true, because the fourth matrix has a missing value. SYSTAT uses the EM algorithm to reestimate this value, compute a new metric solution, and iterate two more times until convergence. This extra set of iterations did not do much for you in

this example because the stress is insignificantly higher than it would have been had you stopped at only two iterations. With many missing values, however, the EM algorithm will improve MDS solutions substantially.

For the INDSCAL model, you have a set of coordinates for the colas and one for the subjects. In the three-dimensional graph of the coordinates, the colas are represented by symbols and the subjects by vectors. The first dimension separates the diet colas from the others. The second dimension differentiates between Dr. Pepper/diet Dr. Pepper and the remaining colas.

For each subject, you have a contribution to overall stress and a separate squared correlation (RSQ) between the predicted and obtained distances in the configuration. Notice that subject 10 is fit worst (STRESS = 0.211658) and subject 9 best (STRESS = 0.163263). Furthermore, subjects 1, 5, and 6 have a high loading on the first dimension, indicating that they place a higher emphasis on diet/nondiet differences than on cherry cola/cola differences. Subjects 7, 8, and 10, on the other hand, emphasize the second dimension more.

#### **Example 4** **Nonmetric Unfolding**

The *COLRPREF* data set contains color preferences among 15 SYSTAT employees for five primary colors. This example uses the MDS unfolding model to scale the people and the colors in two dimensions, such that each person's coordinate is near his or her favorite color's coordinate and far from his or her least favorite color's coordinate. For this example, use ROWS to specify the number of rows for a rectangular matrix and SHAPE to specify the type of matrix input to use. When you enter these data for the first time, you must remember to specify their type as DISSIMILARITY so that small numbers are understood as meaning most similar (preferred).

To scale these with the unfolding model, specify:

```
MDS
  USE COLRPREF
  MODEL red .. blue / SHAPE=RECT
  IDVAR name$
  ESTIMATE / SPLIT=ROWS
```

Notice that you are using the Kruskal loss function as the default.

## The output is:

Monotonic Multidimensional Scaling  
 Kruskal Method  
 The data are analyzed as dissimilarities  
 The data are rectangular (lower corner matrix)  
 Fitting is split between rows of data matrix  
 Minimizing Kruskal STRESS (form 1) in 2 dimensions

## Iteration History

Iteration	STRESS
0	0.148373
1	0.135423
2	0.125152
3	0.117255
4	0.111131
5	0.106394
6	0.102622
7	0.099539
8	0.096883
9	0.094498
0	0.107455
1	0.100496
2	0.096037
3	0.092747
4	0.090087

Stress of Final Configuration : 0.090087  
 Proportion of Variance (RSQ) : 0.940008

## Coordinates in 2 Dimensions

Variable	Dimension	
	1	2
RED	0.252839	-0.486827
ORANGE	0.530030	-1.697840
YELLOW	-1.312679	-0.563914
GREEN	1.388778	0.255362
BLUE	-0.548163	0.785062
Patrick	0.560619	0.782517
Laszlo	-0.728868	-0.132010
Mary	-1.005806	0.113803
Jenna	0.194159	-0.247226
Julie	-0.702923	-0.219116
Steve	1.176419	-0.756052
Phil	0.612587	0.614672
Mike	-0.802781	-0.017760
Keith	0.273582	0.758853
Kathy	0.048997	0.756548
Leah	-0.718963	0.004649
Stephanie	0.498464	0.577649
Lisa	0.784008	0.209336
Mark	-0.565003	0.500289
John	0.064703	-1.237996

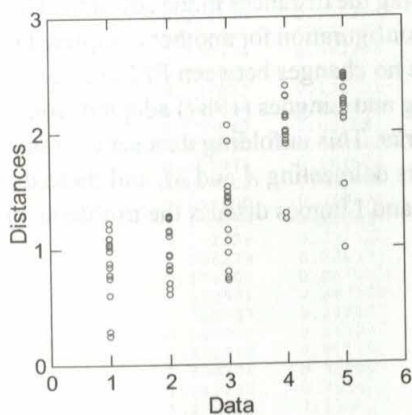
## Row Fit Measures

Row	Stress	RSQ
Patrick	0.000000	1.000000
Laszlo	0.068318	0.969913
Mary	0.004396	0.999893
Jenna	0.048405	0.983337
Julie	0.271710	0.508263

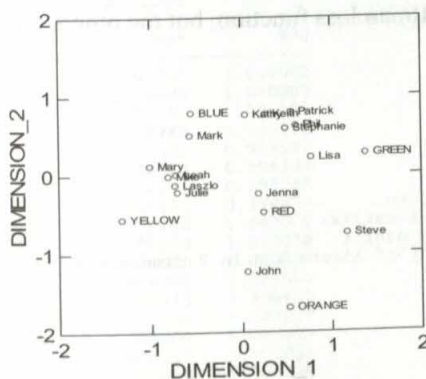


Steve	0.033042	0.992776
Phil	0.061234	0.972002
Mike	0.083004	0.958462
Keith	0.171898	0.773657
Kathy	0.000000	1.000000
Leah	0.067386	0.971396
Stephanie	0.028564	0.993661
Lisa	0.055084	0.980702
Mark	0.000000	1.000000
John	0.024703	0.996053

Shepard Diagram



Configuration



### *Nonmetric Unfolding and the EM Algorithm*

The nonmetric unfolding model has often presented problems to MDS programs because so much data are missing. If you think of the unfolding matrix as the lower corner matrix in a larger triangular matrix of subjects + objects, you can visualize how much data (namely, all of the subject-object comparisons) are missing. Since SYSTAT uses the EM algorithm for missing values, unfolding models do not degenerate as frequently. SYSTAT does a complete MDS using all available data and then estimates missing dissimilarities/similarities using the distances in the solution. These estimated values are then used to get a starting configuration for another complete iteration cycle. This process continues until there are no changes between EM cycles.

The following example, from Borg and Lingoes (1987) adapted from Green and Carmone (1970), shows how this works. This unfolding data set contains dissimilarities only between the points delineating *A* and *M*, and these dissimilarities are treated only as rank orders. Borg and Lingoes discuss the problems in fitting an unfolding model to these data.

The input is:

```
MDS
  USE AM
  IDVAR row$
  MODEL / SHAPE=RECT
  ESTIMATE / LOSS=GUTTMAN  SPLIT=ROWS
```

Notice that the example uses the Guttman loss function, but the others provide similar results.

The output is:

```
Monotonic Multidimensional Scaling
Guttman loss function
The data are analyzed as dissimilarities
The data are rectangular (lower corner matrix)
Fitting is split between rows of data matrix
Minimizing Guttman/Lingoes Coefficient of Alienation in 2 dimensions
```

#### Iteration History

Iteration	Alienation
0	0.076137
1	0.037826
2	0.023535
3	0.017735
4	0.013271
5	0.009960

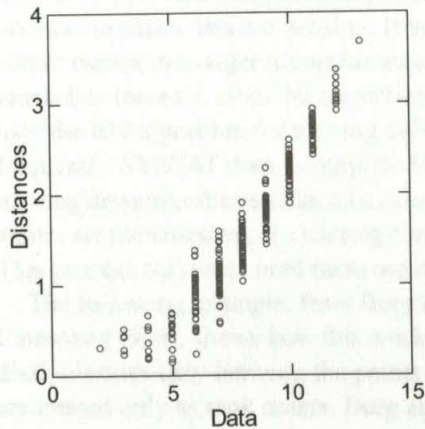
Alienation of Final Configuration : 0.009960  
 Proportion of Variance (RSQ) : 0.999247  
 Coordinates in 2 Dimensions

Variable	Dimension	
	1	2
A1	-0.938673	-1.018145
A2	-0.892414	-0.975977
A3	-1.090552	-0.414280
A4	-1.066410	-0.398294
A5	-1.187946	0.146240
A6	-1.227090	0.337007
A7	-1.543054	0.668773
A8	-0.997198	0.552347
A9	-0.694101	0.467134
A10	-0.305124	0.356277
A11	0.014600	0.102324
A12	0.104769	0.102859
A13	0.130734	0.092203
A14	-0.845901	0.094247
A15	-0.739913	0.136811
A16	-0.569064	0.128649
M1	0.735047	-1.080081
M2	0.430679	-0.524410
M3	0.201071	-0.564505
M4	0.013212	-0.431126
M5	-0.154900	-0.326271
M6	-0.205833	-0.180667
M7	-0.172336	0.121768
M8	-0.056279	0.224731
M9	0.175900	0.267054
M10	0.560531	0.243136
M11	0.588937	0.218047
M12	0.588937	0.218047
M13	0.831710	0.871193
M14	0.890298	0.660027
M15	1.041142	0.212429
M16	1.238422	0.156627
M17	1.498853	0.231883
M18	1.701128	-0.210182
M19	1.940814	-0.485875

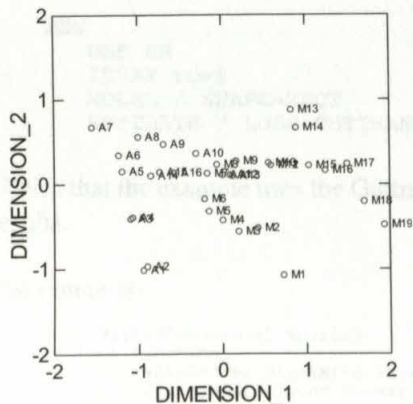
## Row Fit Measures

Row	Stress	RSQ
M1	0.000000	1.000000
M2	0.000000	1.000000
M3	0.000000	1.000000
M4	0.000463	0.999998
M5	0.027442	0.993181
M6	0.022243	0.996393
M7	0.024279	0.997286
M8	0.016153	0.998870
M9	0.000250	1.000000
M10	0.000000	1.000000
M11	0.000000	1.000000
M12	0.000000	1.000000
M13	0.001745	0.999957
M14	0.000000	1.000000
M15	0.000000	1.000000
M16	0.000000	1.000000
M17	0.000000	1.000000
M18	0.000000	1.000000
M19	0.000000	1.000000

Shepard Diagram



Configuration



### Example 5 Power Scaling Ratio Data

As similarities or dissimilarities are often collected as rank-order data, the nonmetric MDS model has to work “backward” in order to solve for a configuration fitting the data. As J. D. Carroll has pointed out, the MDS model should really express observed data as a function of distances between points in a configuration rather than the other



way around. If your data are direct or derived distances, however, you should try setting `REGRESSION = POWER` with `LOSS = KRUSKAL`. This way, you can fit a Stevens power function to the data using distances between points in the configuration. The results may not always differ much from nonmetric or linear or log MDS, but SYSTAT will also tell you the exponent of the power function in the Shepard diagram. Notice, with this model, that the data and distances are transposed in the Shepard diagram because loss is being computed from errors in the data rather than the distances. SYSTAT calls the loss for the power model `PSTRESS` to distinguish it from Kruskal's `STRESS`. In `PSTRESS`, you use `DATA` and its `DHAT` instead of `DIST` to compute the loss.

The *HELM* data set contains highly accurate estimates of distance between color pairs by one experimental subject (CB). These are from Helm (1959) and reprinted by Borg and Lingoes (1987).

To scale these with power model, specify:

```
MDS
  USE HELM
  MODEL a .. s
  ESTIMATE / REGRESS=POWER
```

The output is:

Power regression function, where  $\text{Dissimilarities} = a * \text{Distances}^p$   
 Kruskal Method  
 The data are analyzed as dissimilarities  
 Minimizing `PSTRESS` (`STRESS` with `DIST` and `DATA` exchanged) in 2 dimensions

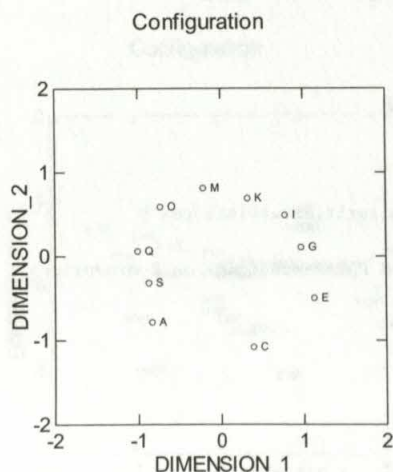
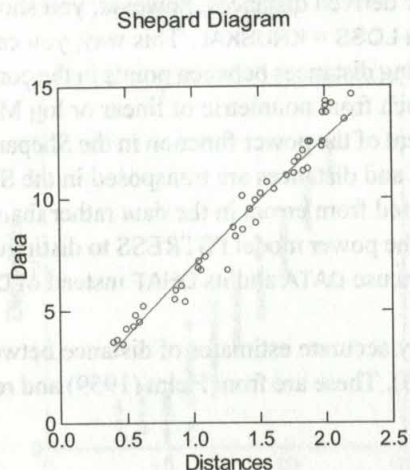
Iteration History

Iteration	PSTRESS
0	0.142062
1	0.131426
2	0.127137
3	0.125205

Stress of Final Configuration	: 0.125205
Estimated Exponent for Power Regression	: 0.851539
Proportion of Variance (RSQ)	: 0.910392

Coordinates in 2 Dimensions

Variable	Dimension	
	1	2
A	-0.828615	-0.792411
C	0.396618	-1.087634
E	1.134571	-0.503104
G	0.977829	0.101019
I	0.785506	0.483283
K	0.331216	0.683545
M	-0.205344	0.804234
O	-0.725019	0.581419
Q	-0.999584	0.052736
S	-0.867177	-0.323088



SYSTAT estimated the power exponent for the function, fitting distances to dissimilarities as 0.85. Color and many other visual judgments show similar power exponents less than 1.0.

## Computation

This section summarizes algorithms separately for the Kruskal and Guttman methods. The algorithms in these options substantially follow those of Kruskal (1964a, 1964b)

and Guttman (1968). MDS output should agree with other nonmetric multidimensional scaling except for rotation, dilation, and translation of the configuration. Secondary documentation can be found in Schiffman, Reynolds, and Young (1981) and the other multidimensional scaling references. The summary assumes that dissimilarities are input. If similarities are input, MDS inverts them.

## Algorithms

### Kruskal Method

The program begins by generating a configuration of points whose interpoint distances are a linear function of the input data. For this estimation, MDS uses a metric multidimensional scaling. Missing values in the input dissimilarities matrix are replaced by mean values for the whole matrix. Then the values are converted to distances by adding a constant. A scalar products matrix  $\mathbf{B}$  is then calculated following the procedures described in Torgerson (1958). The initial configuration matrix  $\mathbf{X}$  in  $p$  dimensions is computed from the first  $p$  eigenvectors of  $\mathbf{B}$  using the Young-Householder procedure (Torgerson, 1958).

After an initial configuration is computed by the metric method, nonmetric optimization begins (there are no metric pre-iterations). At the beginning of each iteration, the configuration is normalized to have zero centroid and unit dispersion. Next, Kruskal's DHAT (fitted) distance values are computed by a monotonic regression of distances onto data. Tied data values are ordered according to their corresponding distances in the configuration.

Stress (formula 1) is calculated from fitted distances, observed distances, and input data values. If the stress is less than 0.001, or has decreased less than 0.001 per iteration in the last five iterations, or the number of iterations equals the number specified by the user (default is 50), iterations terminate (that is, go to the next paragraph). Otherwise, the negative gradient is computed for each point in the configuration by taking the partial derivatives of stress with respect to each dimension. Points in the configuration are moved along their gradients with a step size chosen as a function of the rate of descent; the steeper the descent, the smaller the step size. This completes an iteration.

After the last iteration, the configuration is shifted so that the origin lies in the centroid. Thus, the point coordinates sum to 0 on each dimension. Moreover, the configuration is normalized to unit size so that the sum of squares of its coordinates is 1. If the Minkowski constant is 2 (Euclidean scaling, which is the standard option), the final configuration is rotated to its principal axis.



### **Guttman Method**

The initial configuration for the Guttman option is computed according to Lingoes and Roskam (1973). Principal components are computed on a matrix  $C$ ,

$$c_{ij} = 1 - \frac{r_{ij}}{\frac{n(n-1)}{2}}$$

where  $r_{ij}$  are the ranks of the input dissimilarities (smallest rank corresponding to smallest dissimilarity), and  $n$  is the number of points. The diagonal elements of  $C$  are

$$c_{jj} = 1 - \sum r_{ij}$$

where the sum is taken over the entire row of the dissimilarity matrix.

For the iteration stage, the initial configuration is normalized as in the Kruskal method. Then rank images corresponding to each distance in the configuration are computed by permuting the configuration distances so that they mirror the rank order of the original input dissimilarities. Ties in the data are handled as in the Kruskal method. These rank images are used to compute the Guttman/Lingoes coefficient of alienation. Iterations are terminated if this coefficient becomes arbitrarily small, if the number of iterations exceeds the maximum, or if the change in its value becomes small. Otherwise, the points in the configuration are moved five times using the same rank images but different interpoint distances each time to compute a new negative gradient. These five cycles within each iteration are what lengthens the calculations in the Guttman method. This completes an iteration.

The final configuration is rotated and scaled as with the Kruskal method. Guttman/Lingoes programs normalize the extreme values of the configuration to unity and thus do not plot the configuration with a zero centroid, so MDS output corresponds to their output within rigid motion and configuration size.

### **Missing Data**

Missing values in a similarity/dissimilarity matrix are ignored in the computation of the loss function that determines how points in the configuration are moved. For information on how this function is computed, see the discussion of algorithms.



If you compute a similarity matrix with Correlations for input to MDS, the matrix will have no missing values unless all of your cases in the raw data have a constant or missing value on one or more variables.

## References

- \* Borg, I. and Groenen, P. (2005). *Modern multidimensional scaling: Theory and applications*. 2nd ed. New York: Springer-Verlag.
- Borg, I. and Lingoes, J. (1981). *Multidimensional data representations: When and why?* Ann Arbor: Mathesis Press.
- Borg, I. and Lingoes, J. (1987). *Multidimensional similarity structure analysis*. New York: Springer Verlag.
- Carroll, J. D. and Arabie, P. (1980). Multidimensional scaling. M. R. Rosenzweig and L. W. Porter, eds. *Annual Review of Psychology*, 31, 607–649.
- Carroll, J. D. and Chang, J. J. (1970). Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition. *Psychometrika*, 35, 283–319.
- \* Carroll, J. D. and Wish, M. (1974). Models and methods for three-way multidimensional scaling. D. H. Krantz, R. C. Atkinson, R. D. Luce, and P. Suppes, eds. *Contemporary Developments in Mathematical Psychology, Vol. II: Measurement, Psychophysics, and Neural Information Processing*. San Francisco: W. H. Freeman and Company.
- \* Coombs, C. H. (1964). *A theory of data*. New York: John Wiley & Sons.
- Davison, M. L. (1983). *Multidimensional scaling*. New York: John Wiley & Sons.
- Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology*, 38, 467–474.
- Green, P. E. and Carmone, F. J. (1970). *Multidimensional scaling and related techniques*. Boston: Allyn and Bacon.
- Green, P. E. and Rao, V. R. (1972). *Applied multidimensional scaling*. New York: Holt, Rinehart, and Winston.
- Guttman, L. (1954). A new approach to factor analysis: The radex. P. F. Lazarsfeld, ed. *Mathematical Thinking in the Social Sciences*. New York: Free Press.
- Guttman, L. (1968). A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrika*, 33, 469–506.
- Helm, C. E. (1959). A multidimensional ratio scaling analysis of color relations. *Technical Report*, Princeton University and Educational Testing Service, June 1959.
- Kruskal, J. B. (1964a). Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29, 1–27.
- Kruskal, J. B. (1964b). Nonmetric multidimensional scaling: A numerical method. *Psychometrika*, 29, 115–129.

- Kruskal, J. B., Wish, M and Uslaner, E.M. (2006). *Multidimensional scaling*. Beverly Hills, Calif.: Sage Publications.
- Lingoes, J. C. and Roskam, E. E. (1973). A mathematical and empirical study of two multidimensional scaling algorithms. *Psychometrika Monograph Supplement*, 19.
- Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired-associate learning tasks. *Journal of Experimental Psychology*, 53, 94–101.
- Schiffman, S. S., Reynolds, M. L., and Young, F. W. (1981). *Introduction to multidimensional scaling: Theory, methods, and applications*. New York: Academic Press.
- Shepard, R. N. (1963). Analysis of proximities as a study of information processing in man. *Human Factors*, 5, 33–48.
- Shepard, R. N., Romney, A. K., and Nerlove, S., eds. (1972). *Multidimensional scaling: Theory and application in the behavioral sciences*. New York: Academic Press.
- \* Takane, Y., Young, F. W., and de Leeuw, J. (1977). Nonmetric individual differences scaling: An alternating least squares method with optimal scaling features. *Psychometrika*, 42, 3–27.
- Torgerson, W. S. (1958). *Theory and methods of scaling*. New York: John Wiley & Sons.
- Weinberg, S. L. and Menil, V. C. (1993). The recovery of structure in linear and ordinal data: INDSCAL and ALSCAL. *Multivariate Behavioral Research*, 28:2, 215–233.

(\* indicates additional references.)

# ***Multinormal Tests***

*Mangalmurti Badgajar*

Just as normality plays a vital role in many univariate statistical procedures, multivariate normality plays a crucial role in multivariate data analysis. Results are often obtained after assuming the underlying distribution to be normal; but these results are valid and correct only if the normality assumption is itself justified.

MNTEST assesses the marginal normality of each variable in multivariate data. The Shapiro-Wilk test is used if the sample size is less than or equal to 5000; otherwise, the Lilliefors test (Kolmogorov-Smirnov test with estimated parameters) is favored. MNTEST computes Mardia's skewness and kurtosis coefficients (Mardia, 1970), and performs tests of significance of these coefficients using asymptotic distributions. These tests are generally effective for testing multivariate normality (Mecklin and Mundform, 2004). MNTEST also computes the Henze-Zirkler test statistic (Henze and Zirkler, 1990; Mecklin and Mundform, 2004, also list it amongst potentially useful tests), and the associated *p-value* using the lognormal distribution. Finally, it produces the beta Q-Q plot of scaled squared Mahalanobis distances following the approach of Gnanadesikan and Kettenring (1972).

## ***Statistical Background***

Mardia (1970) has listed some measures of skewness and kurtosis and their distributional properties. Rejection of normality using Mardia's tests indicates that either multivariate outliers are present or the multivariate normal distribution does not describe the data suitably. In addition to the Mardia measures and tests, we also calculate the Henze-Zirkler test statistic (Henze and Zirkler, 1990), which has better



power properties than the Mardia test against symmetric alternatives such as a family of elliptically contoured distributions.

Q-Q plots using 'Squared Mahalanobis Distances' are useful to identify departures from multivariate normality and outliers. Graphical tests alone are inadequate; some numeric measures are needed. Romeu and Ozturk (1993) investigated ten tests of goodness-of-fit for multivariate normality. Their simulation study shows that the multivariate tests of skewness and kurtosis proposed by Mardia (1970) "are the most stable and reliable for assessing multivariate normality" (Timm, 2002: page 121). For evaluating marginal normality we use the univariate Shapiro-Wilk test (Shapiro and Wilk, 1965).

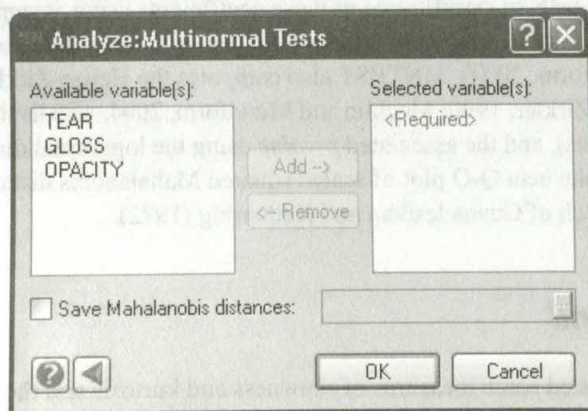
## Multinormal Tests in SYSTAT

### Multinormal Tests Dialog Box

To open the Multinormal Tests dialog box, from the menus choose:

Analyze

Multinormal Tests...



**Selected variables.** Select two or more numeric variables for testing multivariate normality.

**Save Mahalanobis distances.** Saves data and squared Mahalanobis distances.



## Using Commands

First, specify your data with *USE filename*. Continue with:

```
SSAVE filename / MAHAL  
MNTEST varlist
```

MNTEST assesses the marginal normality for each variable in *varlist*, using the Shapiro-Wilk test, if sample size is less than or equal to 5000; otherwise, it uses the Lilliefors test (Kolmogorov-Smirnov test with estimated parameters). Further, it computes Mardia's skewness and kurtosis coefficients for the variables in *varlist* and performs a test of the significance of these coefficients using an asymptotic distribution. The Henze-Zirkler test statistic and its associated *p-value* using lognormal distribution are also displayed. Finally, it produces the beta Q-Q plot of scaled squared Mahalanobis distances.

SSAVE with MAHAL option will save data and squared Mahalanobis distances in the file *filename*.

## Usage Considerations

**Type of data.** MNTEST uses rectangular numeric data.

**Print options.** The output is standard for all PLENGTH options.

**Quick Graphs.** MNTEST produces a Q-Q plot of scaled squared Mahalanobis distances.

**Saving files.** MNTEST saves data and squared Mahalanobis distances.

**BY groups.** MNTEST produces a separate output for each group.

**Case frequencies.** FREQUENCY is not available in MNTEST.

**Case weights.** WEIGHT is not available in MNTEST.

## Examples

### Example 1

#### Multivariate Normality Assessment of Perspiration Measurements

In this example we check the multivariate normality for *SWEAT* data from Johnson and Wichern (2002). The data set contains perspiration measurements from twenty healthy females arranged in three components, *SWEAT\_RATE* = sweat rate, *SODIUM* = sodium content, and *POTASSIUM* = potassium content.

The input is:

```
USE SWEAT
MNTTEST SWEAT_RATE SODIUM POTASSIUM
```

The output is:

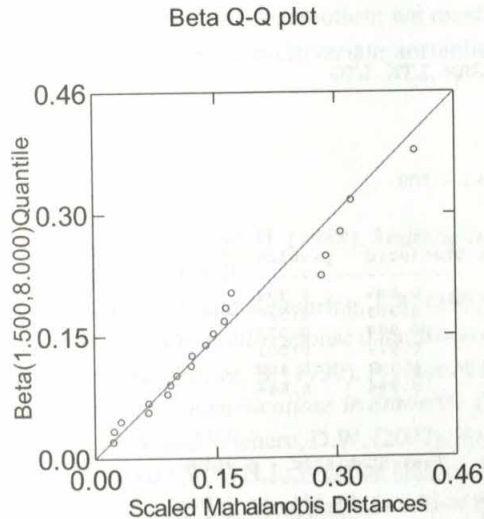
Number of Cases Used for Analysis = 20

##### Marginal Normality Tests

Variable	Test	Test Statistic	p-value
SWEAT_RATE	Shapiro-Wilk	0.976	0.869
SODIUM	Shapiro-Wilk	0.986	0.986
POTASSIUM	Shapiro-Wilk	0.964	0.623

##### Joint Normality

Test	Coefficients	Test Statistic	p-value
Mardia Skewness	2.188	9.033	0.529
Mardia Kurtosis	11.881	-1.273	0.203
Henze-Zirkler		0.452	0.387



By using *p-values* for marginal Shapiro-Wilk test statistics, we get sufficient evidence for marginal normality of variables *SWEAT\_RATE*, *SODIUM*, and *POTASSIUM*. The joint multivariate normality of *SWEAT\_RATE*, *SODIUM*, and *POTASSIUM* is also supported by *p-values* associated with Mardia's skewness, kurtosis coefficients, and the Henze-Zirkler test.

### Example 2

#### *Multivariate Normality Assessment of Anthropometric Measurements*

Here we check the multivariate normality for six variables measured on a selected sample of Swiss army personnel. The variables, as described in Flury and Riedwyl (1988) are:

*MFB* = minimal frontal breadth

*BAM* = breadth of angulus mandibulae

*TFH* = true facial height

*LGAN* = length from glabella to apex nasi

*LTN* = length from tragon to nasion

*LTG* = length from tragon to gnathion

Measurements are made on 200 twenty-year old male soldiers.

The input is:

```
USE HEADDIM
MNTEST MFB BAM TFH LGAN LTN LTG
```

The output is:

Number of Cases Used for Analysis = 200

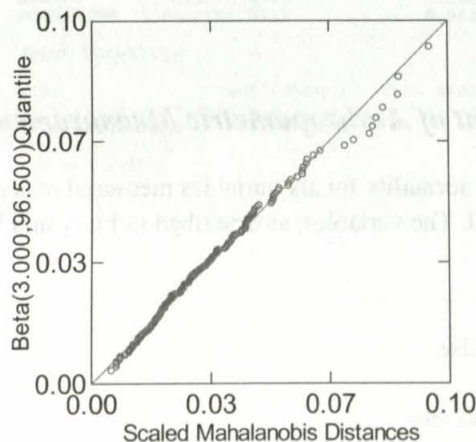
#### Marginal Normality Tests

Variable	Test	Test Statistic	p-value
MFB	Shapiro-Wilk	0.995	0.797
BAM	Shapiro-Wilk	0.993	0.526
TFH	Shapiro-Wilk	0.988	0.078
LGAN	Shapiro-Wilk	0.973	0.001
LTN	Shapiro-Wilk	0.993	0.469
LTG	Shapiro-Wilk	0.994	0.646

#### Joint Normality

Test	Coefficients	Test Statistic	p-value
Mardia Skewness	2.646	89.894	0.003
Mardia Kurtosis	46.939	-0.766	0.444
Henze-Zirkler		0.997	0.140

Beta Q-Q plot



In this case the *p-value* associated with the Shapiro-Wilk test statistic of the variable *LGAN* is very low. Also, the *p-value* for significance testing of Mardia's skewness coefficient is low (0.003). Thus there is no clear-cut evidence that this data set follows



a multivariate normal distribution; we must therefore be careful and cautious while analyzing it under the multivariate normality assumption.

## References

- Flury, B. and Riedwyl, H. (1988). *Multivariate statistics: A practical approach*. London: Chapman and Hall.
- Gnanadesikan, R. and Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multi-response data. *Biometrics*, 28, 81-124.
- Henze, N. and Zirkler, B. (1990). A class of invariant consistent tests for multivariate normality. *Communications in Statistics Theory and Methods*, 19, 3595-3618.
- Johnson, R.A. and Wichern, D.W. (2002). *Applied multivariate statistical analysis*, 5th ed. Englewood Cliffs, N.J.: Prentice Hall.
- Mardia, K. V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, 58, 519-530.
- Mecklin, C. J. and Mundform, D. J. (2004). An appraisal and bibliography of tests for multivariate normality. *International Statistical Review*, 72:1, 123-138.
- Romeu, J. L. and Ozturk, A. (1993). A comparative study of goodness-of-fit tests for multivariate normality. *Journal of Multivariate Analysis*, 46,309-334.
- Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complex samples). *Biometrika*, 52, 591-611.
- Timm, N. H.(2002). *Applied multivariate analysis*. New York: Springer-Verlag.



# Multivariate Analysis of Variance

Sayyad Nisar Badashah and Rajesh V. Nath

(Some material has been taken from the SYSTAT 10.2 manual, Statistics I: Chapter 16: Linear Models III: General Linear Models by Leland Wilkinson and Mark Coward.)

The Multivariate Analysis of Variance (MANOVA) feature handles estimation and testing in one-way, two-way, and multi-way classified multivariate data, repeated measures analysis, and more generally handles within-group and between-group testing. These include multivariate analysis of data obtained by using standard experimental designs and standard factorial treatment structures with crossing and nesting.

You can select any of the three types of sum of squares, Type I, Type II, and Type III, for the analysis. MANOVA begins with a preliminary analysis that provides parameter estimates and least-squares mean vectors. This is followed by results of tests of hypotheses, where, besides results of multivariate tests in terms of suitable statistics and their *p-values*, results of corresponding univariate tests for each (dependent) variable (components of the multivariate data vector) are also provided. AIC, AIC (Corrected) and Schwarz's BIC values are also provided for each fitted model. For more information on AIC and Schwarz's BIC in SYSTAT refer to the Chapter Linear Models: Introduction: "Variable Selection" on page 15 in *Statistics II*.

Resampling procedures are available in this feature.

## Statistical Background

Multivariate Analysis of Variance (MANOVA) is the multivariate analog of the Analysis of Variance (ANOVA). MANOVA procedures were already available in SYSTAT's earlier versions and could be used by invoking the General Linear Model (GLM) procedures and suitably defining the models and the hypotheses, through either dialog or commands. However, many applications of MANOVA are in standard problems, and, in this MANOVA feature, such standard applications have been made simpler by making them menu-driven.

As with ANOVA, the independent variables in a MANOVA model are factors, each factor having two or more levels. Unlike ANOVA, MANOVA deals with multiple dependent variables, rather than a single dependent variable. MANOVA examines whether the population means on a set of dependent variables vary across levels of a factor or factors. For example, suppose three varieties of peanuts were grown at different geographical locations (1, 2) and three variables of interest were measured:  $X_1$  = yield (plot weight),  $X_2$  = sound mature kernels (weight in grams--maximum of 250 grams), and  $X_3$  = seed size (weight, in grams, of 100 seeds). In this two-factor experiment, the primary objective is to compare location effects, variety effects and their interaction. Clearly, a two-way MANOVA is appropriate in this situation.

In most models for which MANOVA is used, the following assumptions are made:

- The joint distribution of dependent variables is multivariate normal in each level of factor combinations.
- The variances and covariances (variance-covariance matrix) among the dependent variables are the same across all levels of factor combinations.
- The multivariate observations are independently distributed over the observational units.

The main interest in MANOVA is the comparison of mean vectors over factor-level combinations. For many problems, the MANOVA procedure is similar to the ANOVA procedure for the corresponding univariate problem, wherein the sum of squares is replaced by a sum of squares and cross-products (SSCP) matrix. Thus, there is a total SSCP matrix that is decomposed into within-groups, i.e., the error or residual SSCP matrix and between-groups SSCP matrix. Further decomposition is carried out depending on the specific models and the hypotheses being tested. While the test statistic in ANOVA is the ratio of mean squares with an appropriate F-distribution under the hypothesis, in MANOVA, the test statistics are generalized versions of these ratios based upon corresponding SSCP matrices with their sampling distributions often approximated by suitable F distributions.



## MANOVA Tests

SYSTAT provides the following four multivariate test statistics for testing the significance of various effects in a model. The following notations are used to represent various SSCP matrices:

**G:** Within-groups (error) SSCP matrix

**H:** Between-groups SSCP matrix

**T:** Total SSCP matrix

### *Wilks's Lambda ( $\Lambda$ or $W$ or likelihood ratio criterion)*

The first of these four statistics is the Wilks's Lambda:

$$W = |G| / |T|$$

The statistic is a monotonically decreasing function of the log-likelihood ratio statistic. The value of the test statistic varies from 0 to 1. The distribution of  $W$  is approximated by the  $F$  distribution (Rao, 1973).

### *Pillai's Trace ( $V$ )*

The statistic is

$$V = \text{trace}(HT^{-1})$$

An approximate  $F$ -ratio is displayed in SYSTAT.

### ***Hotelling-Lawley Trace (T)***

The statistic is

$$T = \text{trace}(G^{-1}H)$$

The *F*-ratio approximation is similar to that of the other statistics.

### ***Roy's Greatest Root (Theta)***

This statistic is derived by Roy's union-intersection approach to MANOVA. Along with  $\lambda_1$ , the largest eigenvalue of the matrix  $G^{-1}H$ , SYSTAT displays the following form of the statistic:

$$\theta = \frac{\lambda_1}{1 + \lambda_1}$$

The exact values for the probabilities are taken from the Heck (1960) chart. The chart for the percentile points of distribution of the largest root is commonly given for  $\theta$ . This is more powerful than the others if the mean vectors are collinear.

Historically, Wilks's lambda played a dominant role in tests in MANOVA because it was the first to be derived and in view of its flexibility and its well-known *F* approximation. In the case of two groups, all the four test statistics are equivalent and, in turn, equivalent to Hotelling's  $T^2$  statistic. In those cases when these statistics differ with regard to the acceptance or rejection of the null hypothesis, you can examine the eigenvalues to select the best one from the procedures discussed above.

Since, like ANOVA, MANOVA is derived from the GLM module, you can find a detailed discussion of various aspects of estimation and testing in Chapter 1 of "Linear Models" on page 1 of *Statistics II*. Also, for further information, see Johnson and Wichern (2002), Rencher (2002), or Timm (2002).

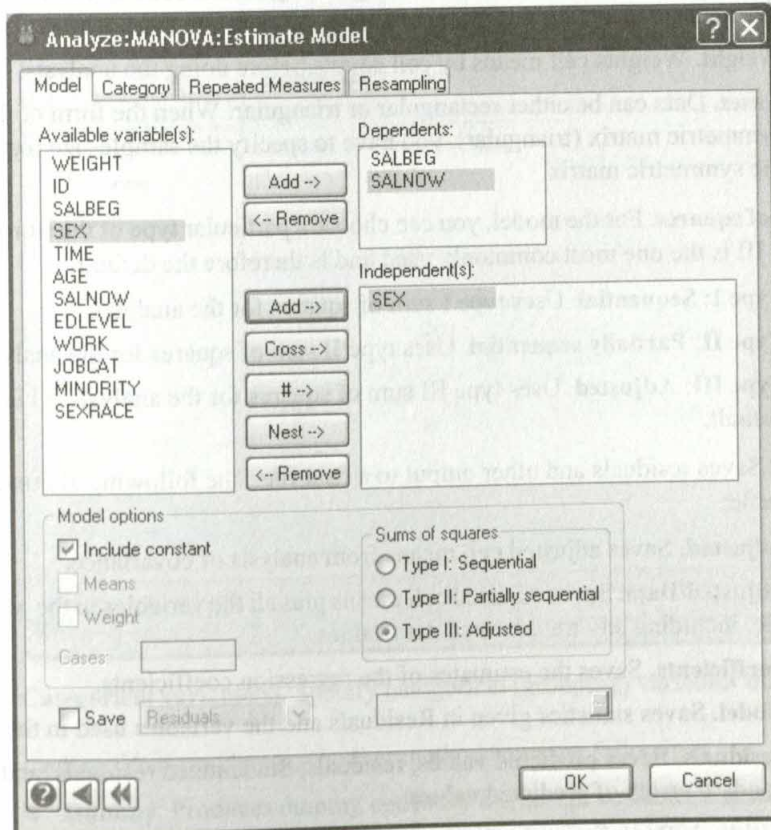
## MANOVA in SYSTAT

### MANOVA: Estimate Model Dialog Box

Estimate Model produces estimates of parameters and tests for equality of group effects.

To open the MANOVA: Estimate Model dialog box from the menus choose:

Analyze  
MANOVA  
Estimate Model....



**Dependents.** Select the response variables you want to examine. The dependent variables should be continuous numeric variables.

**Independent(s).** Select one or more categorical or numerical variables. The variables not specified as categories are treated as covariates. If you want to build a model that contains interaction effects and nested effects, use Cross and Nest buttons to build the model. If you want to include effects like  $A+B+A*B$ , select the effects A and B, and then click on the # button.

**Model options.** The following model options are available:

- **Include constant.** This includes a constant term in your model. Deselect this option to remove the constant.
- **Means.** Specifies a fully factorial design using means coding (For more information on means coding, see “Linear Models” on page 1 in *Statistics II*).
- **Weight.** Weights cell means by cell counts before doing the analysis.
- **Cases.** Data can be either rectangular or triangular. When the form of the data is a symmetric matrix (triangular), you have to specify the sample size that generated the symmetric matrix.

**Sum of squares.** For the model, you can choose a particular type of the sum of squares. Type III is the one most commonly used and is therefore the default.

- **Type I: Sequential.** Uses type I sum of squares for the analysis.
- **Type II: Partially sequential.** Uses type II sum of squares for the analysis.
- **Type III: Adjusted.** Uses type III sum of squares for the analysis. This is the default.

**Save.** Saves residuals and other output to a data file. The following alternatives are available:

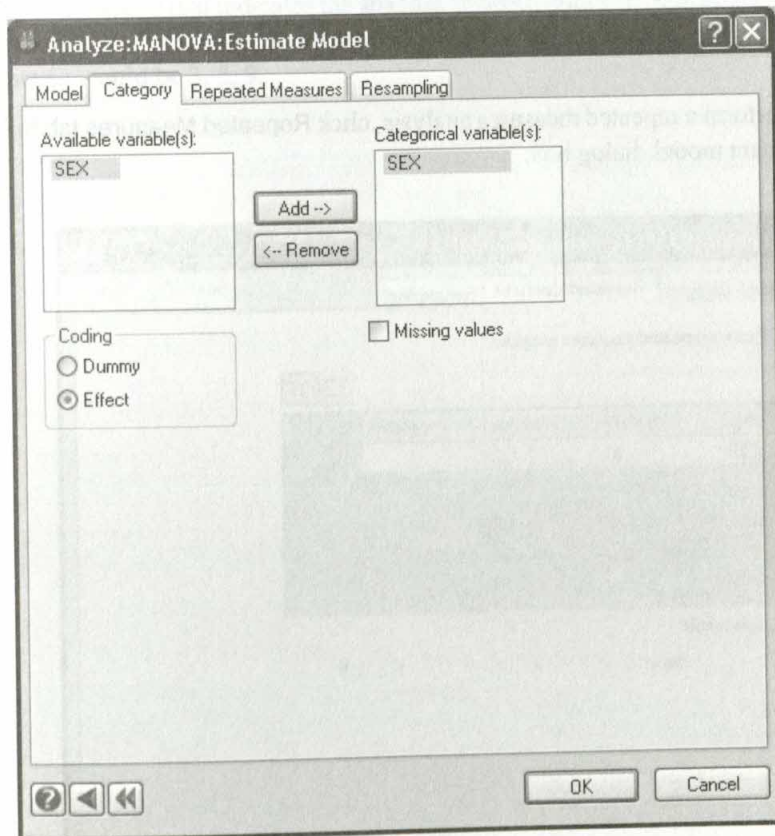
- **Adjusted.** Saves adjusted cell means from analysis of covariance.
- **Adjusted/Data.** Saves adjusted cell means plus all the variables in the working data file, including any transformed data values.
- **Coefficients.** Saves the estimates of the regression coefficients.
- **Model.** Saves statistics given in Residuals and the variables used in the model.
- **Residuals.** Saves predicted values, residuals, Studentized residuals, and the standard errors of predicted values.
- **Residuals/Data.** Saves the statistics given by Residuals, plus all the variables in the working data file, including any transformed data values.



### Category

You can specify numeric or character-valued categorical (grouping) variables that define cells. The variables that are not specified as categorical variables are considered as covariates.

To do so, click the Category tab in MANOVA: Estimate Model dialog box.



**Categorical variable(s).** Specify categorical (grouping) variables that define cells.

**Coding.** You can choose a coding method from the following:

- **Dummy.** Produces dummy codes for the design variables instead of effect codes. Coding of dummy variables is the classic analysis of variance parameterization, in

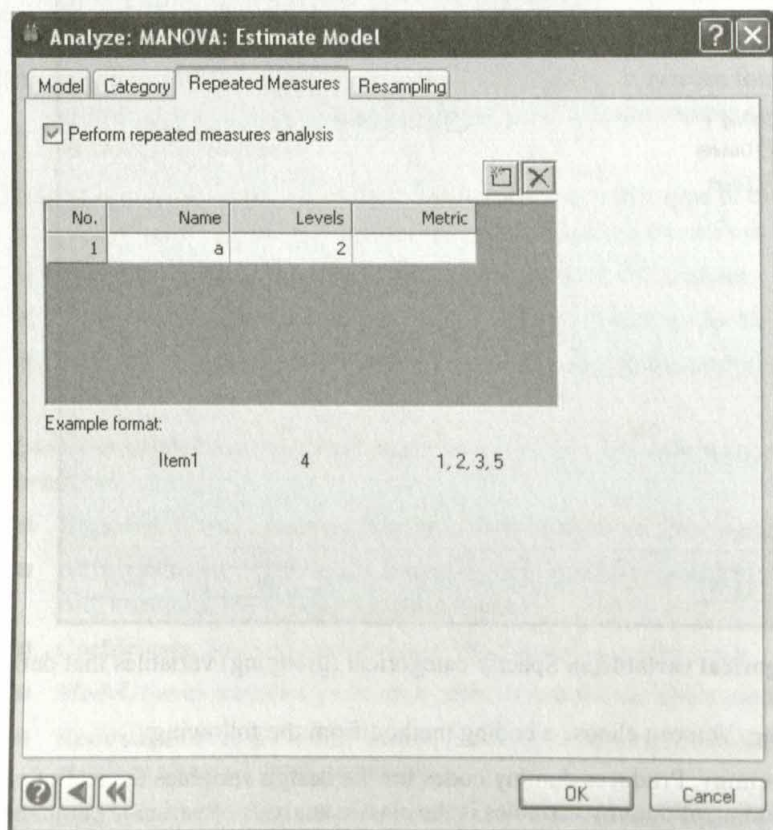
which the sum of effects estimated for a classifying variable is 0. If your categorical variable has  $k$  categories,  $k-1$  dummy variables are created.

- **Effect.** Produces by default, the parameter estimates that are differences from group means.

**Missing values.** Check this to include categorical variables with missing values as a separate category in the analysis.

### *Repeated Measures*

To perform a repeated measures analysis, click Repeated Measures tab in MANOVA: Estimate model dialog box.



**No.** Displays the serial number.

**Name.** Specify names that identify each set of repeated measures.

**Levels.** Enter the number of repeated measures in the set. For example, suppose you have three dependent variables that represent measurements at different times, the number of levels is three.

**Metric.** Metric that indicates the spacing between unevenly spaced measurements. For example, suppose measurements were taken at the third, fifth, and ninth weeks, the metric would be 3, 5, 9.

## Hypothesis Test Dialog Box

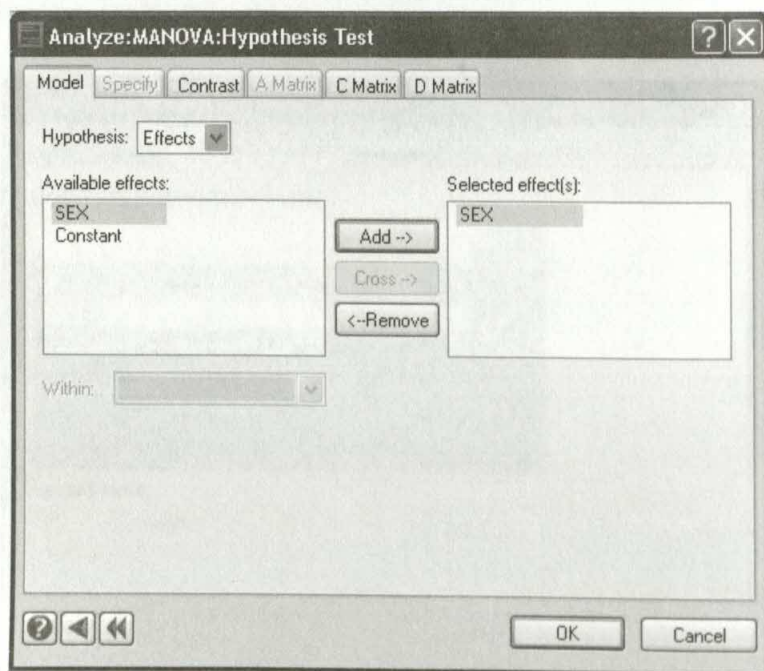
After estimating the treatment effects, contrasts are used to test the relationship among various treatment levels. We may focus on whether the interaction is significant for some linear combination of variables or each variable individually.

To perform the hypothesis tests, from the menus choose:

Analyze

MANOVA

Hypothesis Test...



**Selected effect(s).** Effect or effects selected for testing.

**Hypothesis.** Select the type of hypothesis. The following choices are available:

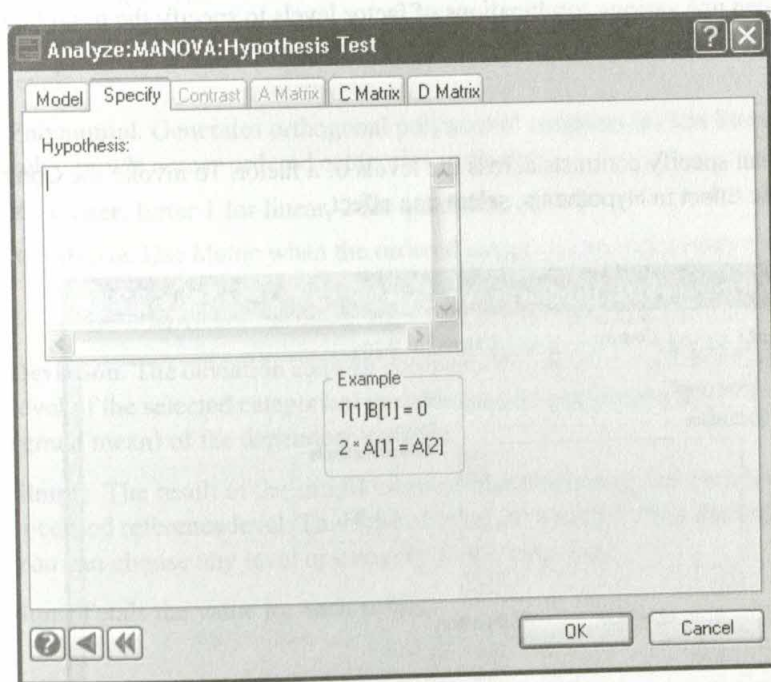
- **Model.** Select to test the significance of the model parameters.
- **Effects.** Select one or more effects you want to test.
- **Specify.** Select to use Specify tab.
- **A Matrix.** Select to use A Matrix tab.



**Within.** Use when specifying a contrast across the levels of repeated measures factor. Select the name assigned to the set of repeated measures in the Repeated Measures tab. This will be enabled only when a repeated measures analysis is performed.

### Specify

To specify the contrasts for between-subjects effects, choose Specify in the MANOVA: Hypothesis Test dialog box.



You can define contrasts across the levels of a grouping variable in a multivariate model. For example, for a two-way factorial MANOVA design with *GENDER*\$ (two categories) and *DRUG* (three categories), you could contrast the marginal mean for the first level of drug against the third level by specifying:

$$DRUG [1] = DRUG [3]$$

Note that the brackets enclose the value of the category (for example, for *GENDER*\$, specify *GENDER\$['MALE']*). For the simple contrast of the first and third levels of *DRUG* for the second *GENDER*\$ only, specify:

```
DRUG[1] GENDER$['MALE'] = DRUG[3] GENDER$['MALE']
```

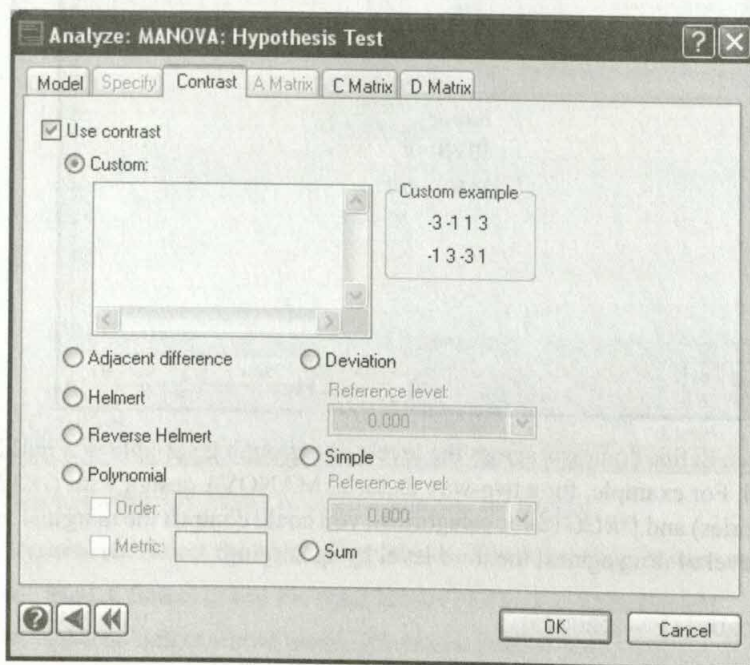
The syntax also allows statements like:

```
-3*DRUG[1] - 1*DRUG[2] + 1*DRUG[3] + 3*DRUG[4]
```

One can use various combinations of factor levels to specify the hypothesis.

### Contrast

You can specify contrasts across the levels of a factor. To invoke the Contrast tab, choose Effect in Hypothesis, select one effect.



Contrast tab generates a contrast for a grouping factor or a repeated measures factor. SYSTAT offers eight types of contrasts:

**Custom.** Enter your own custom coefficients. For example, if your factor has four ordered categories (or levels), you can specify your own coefficients, such as -3 -1 1 3, by typing these values in the Custom text box.

**Adjacent difference.** Compares each level with its adjacent level.

**Helmert.** Compare the mean of each level of the selected factor to the mean of the succeeding levels.

**Reverse Helmert.** Compares the mean of each level of selected factor with the previous levels.

**Polynomial.** Generates orthogonal polynomial contrasts (to test linear, quadratic, or cubic trends across ordered categories or levels).

■ **Order.** Enter 1 for linear, 2 for quadratic, etc.

■ **Metric.** Use Metric when the ordered categories are not evenly spaced. For example, when repeated measures are collected at weeks 2, 4, and 8, enter 2, 4, 8 as the metric.

**Deviation.** The deviation contrast compares the mean of the dependent variable at each level of the selected categorical variable (except a reference level) to the overall mean (grand mean) of the dependent variable.

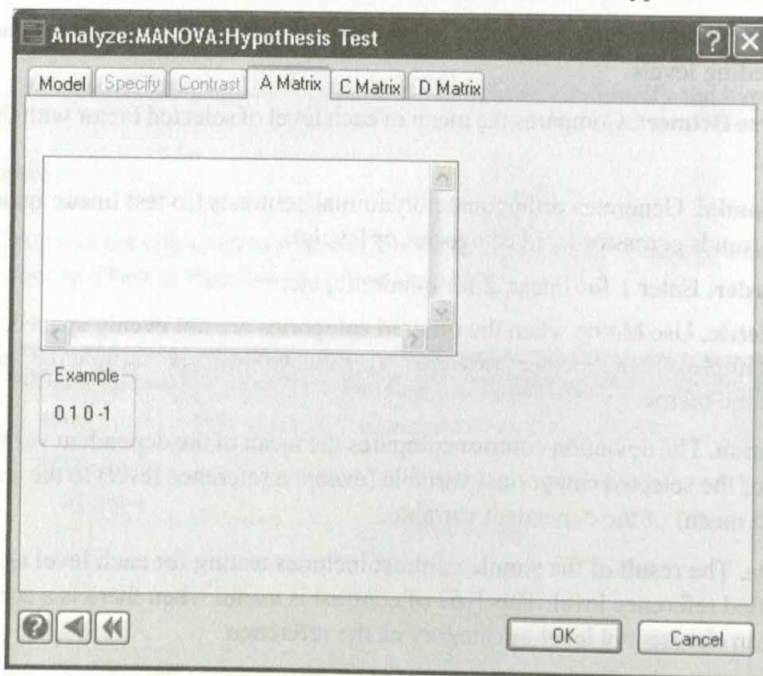
**Simple.** The result of the simple contrast includes testing for each level against the specified reference level. This type of contrast is useful when there is a control group. You can choose any level or category as the reference.

**Sum.** Totals the value for each subject.

### *A, C, and D matrices*

The matrices **A**, **C**, and **D** are available for hypothesis testing in multivariate models. These matrices (**A**, **C**, and **D**) may be specified in several alternative ways; if they are not specified, they assume the default values.

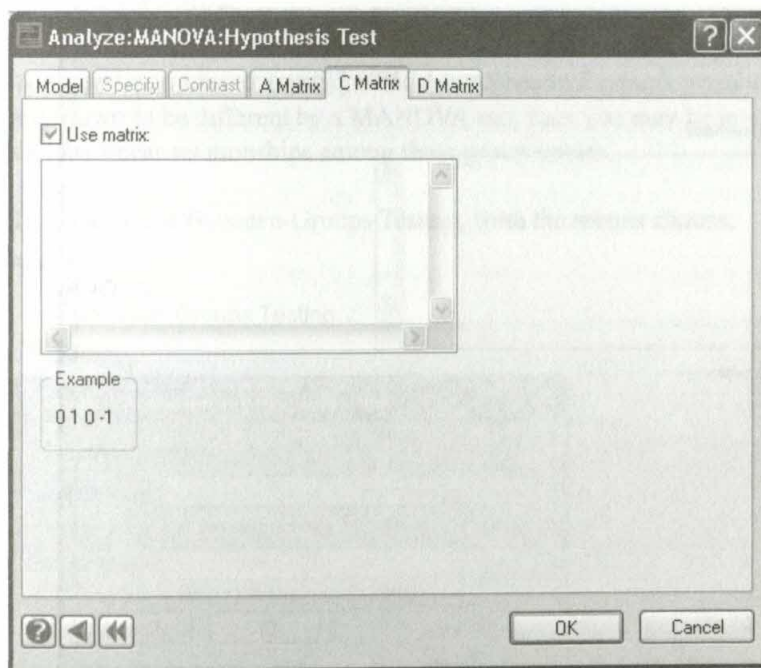
To specify **A** Matrix, click **A matrix** tab in MANOVA: Hypothesis test dialog box.



**A** is a matrix of linear weights contrasting the coefficient estimates (the rows of **B**). You can write your hypothesis in terms of the **A** matrix. The **A** matrix has as many columns as there are regression coefficients (including the constant) in your model. The number of rows in **A** determines how many degrees of freedom your hypothesis involves.

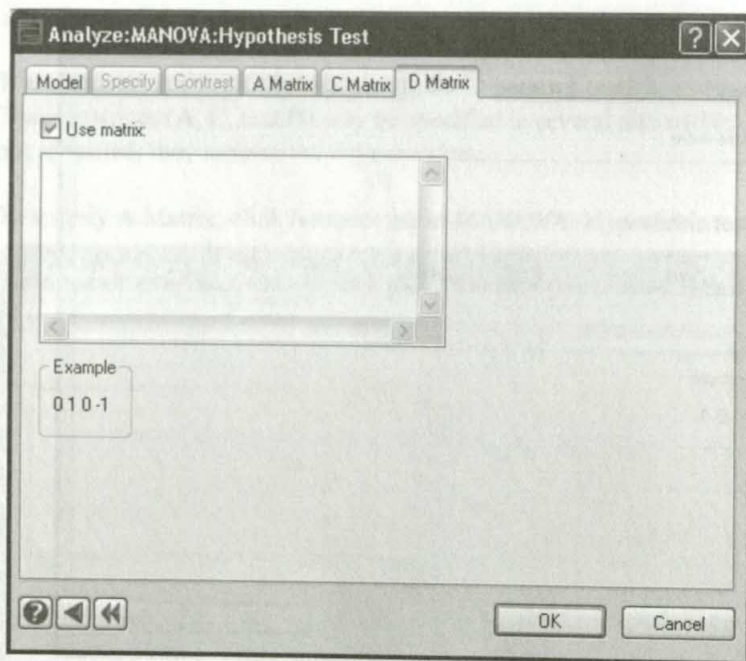
To specify **C** Matrix, click **C Matrix** tab in the MANOVA: Hypothesis Test dialog box.





The **C** matrix is used to test hypotheses for repeated measures analysis of variance designs and models with multiple dependent variables. **C** has as many columns as there are dependent variables. By default, the **C** matrix is the identity matrix.

To specify **D** Matrix, click **D Matrix** tab in the MANOVA: Hypothesis Test dialog box



**D** is a null hypothesis matrix. By default it is a null matrix. The **D** matrix, if you use it, must have the same number of rows as **A**. For univariate multiple regression, **D** has only one column. For multivariate models (multiple dependent variables), the **D** matrix has one column for each dependent variable.

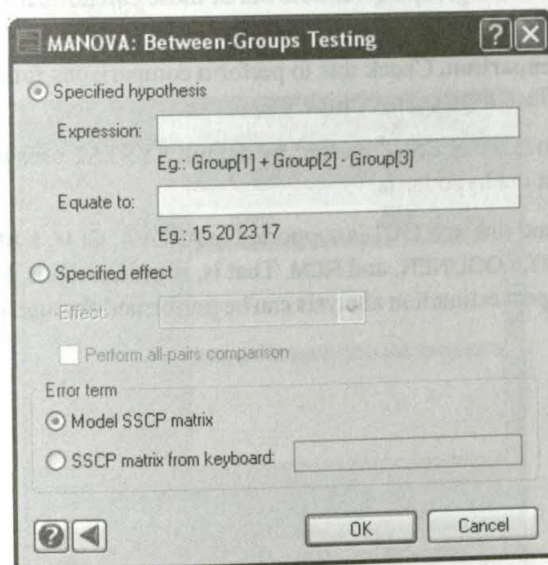
Toggling among command line and GUI is supported in ANOVA, GLM, MANOVA, REGRESS, MIXED, LOGIT, LOGLINER, and RSM. That is, if estimation is performed through dialog box then post estimation analysis can be performed through commands and vice-versa.

## Between-Groups Testing

You may be interested in various linear hypotheses of group means. If the group means are shown to be different by a MANOVA test, then you may be interested in testing various linear relationships among these group means.

To perform the Between-Groups Testing, from the menus choose:

Analyze  
MANOVA  
Between-Groups Testing...



**Specified hypothesis.** Select this option to specify the hypothesis to be tested.

- **Expression.** Enter your expression. For a two-way factorial MANOVA design with *DISEASE* (three categories) and *DRUG* (four categories), you could contrast the group mean for the first level of drug against the third level by specifying:

$DRUG [1] = DRUG [3]$

Alternatively, you can use  $DRUG [1] - DRUG [3]$

Note that the brackets enclose the value of the category (for example, for *GENDER*\$, specify *GENDER\$*['MALE']).

The syntax also allows statements like:

```
- 3*DRUG[1] - 1*DRUG[2] + 1*DRUG[3] + 3*DRUG[4]
```

- **Equate to.** Specify a vector of size equal to the number of dependent variables; if not specified, by default, SYSTAT takes a zero vector of appropriate dimension. 'Equate to' does not allow you to give a group name like *DRUG[2]*; SYSTAT expects that 'Equate to' should be a user-specified numeric vector.

**Specified effect.** Click to perform one-way MANOVA or All Pairs Comparison.

- **Effect.** Shows a list of categorical variables, which have been used to fit the MANOVA model. Select a grouping variable out of those categorical variables to perform a One-Way MANOVA.
- **Perform all pairs comparison.** Check this to perform comparisons for all pairs within a specified effect or grouping variable.

**Error term.** Enter your own error SSCP matrix; by default SYSTAT uses the model error SSCP matrix to test the hypothesis.

Toggling among command line and GUI is supported in ANOVA, GLM, MANOVA, REGRESS, MIXED, LOGIT, LOGLINER, and RSM. That is, if estimation is performed through dialog box then post estimation analysis can be performed through commands and vice-versa.

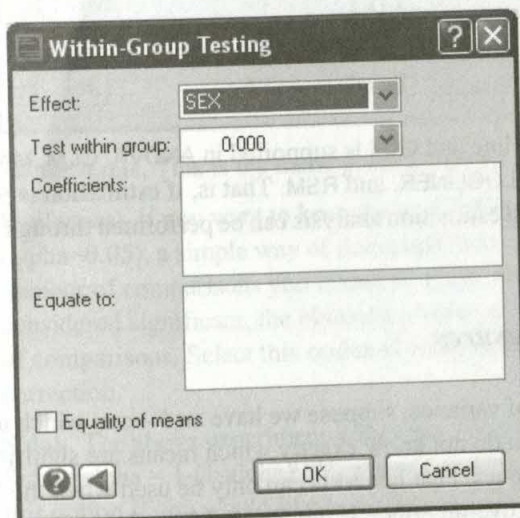


## Within-Group Testing

A hypothesis of interest may be whether there exists a difference in the various components rather than groups. The components are considered inside each group. SYSTAT automatically tests the equality of various components in a group. You can give linear contrasts of your interest; and you can perform the test for equality of component means.

To perform Within-Group Testing, from the menus choose:

Analyze  
MANOVA  
Within-Group Testing...



**Effect.** Effect gives you a list of all categorical variables, which have been used in the model. Select one among them. By default, it takes the first categorical variable in the category list.

**Test within group.** Displays all the levels of the above-specified effect. If you have a categorical variable (say) *CLASS* with two levels (1,2) and if you want to perform testing within level 1, then select 1 in test within. If not selected, SYSTAT takes the first level of the selected effect.

**Coefficients.** Specify the coefficients of the linear hypothesis.

**Equate to.** Specify the null hypothesis vector. If you want to test the mean vector of a level of a group equal to some specific value, then specify a vector of proper order; if not specified, by default, SYSTAT takes the zero vector of appropriate dimension.

**Equality of means.** Check this to perform a test of equality of means of components within a group.

For example:

Grouping variable: CLASS

Test within group: CLASS 1

Hypothesis to test:

$$\mu_1 + \mu_2 + \mu_3 = 5$$

$$\mu_1 - \mu_2 + \mu_3 = 6$$

$$\mu_1 - 2\mu_2 + \mu_3 = 6$$

#### Input

Effect = CLASS

Test within group = 1

Coefficients = 1 1 1

1 -1 1

1 -2 1

Equate to = 5 6 6

Toggling among command line and GUI is supported in ANOVA, GLM, MANOVA, REGRESS, MIXED, LOGIT, LOGLINER, and RSM. That is, if estimation is performed through dialog box then post estimation analysis can be performed through commands and vice-versa.

## Post hoc Test for Repeated measures

After performing analysis of variance, suppose we have an *F-ratio* which tells us that means are not equal; we still do not know exactly which means are significantly different from which other ones. Post hoc tests can only be used when the 'omnibus' ANOVA finds a significant overall effect. If the *F-value* for a factor turns out non-significant, you may not want to go further with the analysis. This protects the post hoc test from being used too liberally.

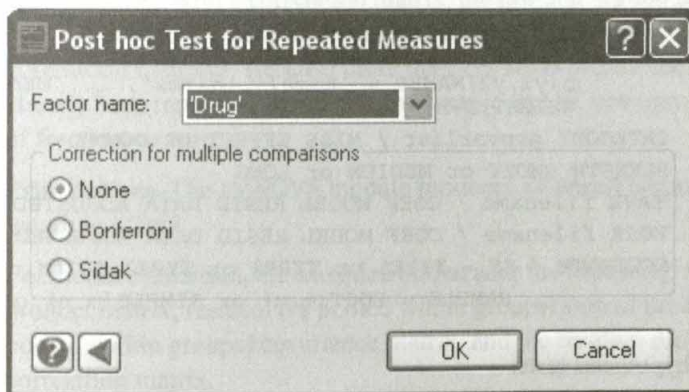
The main problem that designers of post hoc test try to deal with is alpha inflation. This refers to the fact that the more tests you conduct at  $\alpha=0.05$ , the more likely you are to come across a significant difference, which in reality may not exist. The overall chance of a Type I error rate in a particular experiment is referred to as the 'experiment-wise error rate' (or family-wise error rate).

To open the Post hoc Test for Repeated Measures dialog box, from the menus choose:

Analyze

MANOVA

Post Hoc Test for Repeated Measures...



**Factor name.** This is the name given to the set of repeated measures in MANOVA.

**Bonferroni.** If you want to keep the experiment-wise error rate to a specified level ( $\alpha=0.05$ ), a simple way of doing this is to divide the acceptable  $\alpha$  level by the number of comparisons you intend to make. That is, for any one comparison to be considered significant, the obtained  $p$ -value would have to be less than  $\alpha/\text{number of comparisons}$ . Select this option if you would like to perform a Bonferroni correction.

**Sidak.** The above experiment-wise error is kept in control by the use of the formula:  $\text{Sidak\_}\alpha = 1 - (1 - \alpha)^{(1/c)}$ , where  $c$  is the number of paired comparisons. Select this option if you would like to perform a Sidak correction.

Toggling among command line and GUI is supported in ANOVA, GLM, MANOVA, REGRESS, MIXED, LOGIT, LOGLINER, and RSM. That is, if estimation is performed through dialog box then post estimation analysis can be performed through commands and vice-versa.



## Using Commands

Select the data with *USE filename* and continue with:

```
USE filename
MANOVA
  MODEL varlist1 = CONSTANT + varlist2 + var1*var2 +,
        var3(var4)/ Repeat = m, n,..., RPEPEAT m(x1, x2,...),
        n(y1,y2)NAMES = 'name1', 'name2', ..., MEANS,
        WEIGHT,N=n
  CATEGORY grpvarlist / MISS EFFECT OR DUMMY
  PLENGTH SHORT or MEDIUM or LONG
  SAVE filename / COEF MODEL RESID DATA ADJUSTED
  WORK filename / COEF MODEL RESID DATA ADJUSTED
  ESTIMATE / SS = TYPE1 or TYPE2 or TYPE3 QUICK or NOQUICK
        SAMPLE = BOOT(m,n) or SIMPLE(m,n) or JACK
```

To perform hypothesis tests:

```
HYPOTHESIS
EFFECT varlist var1*var2, ...
STANDARDIZE WITHIN or TOTAL
WITHIN 'name'
CONTRAST [matrix] / ADJDIFF or SUM or POLYNOMIAL,ORDER=n,
        METRIC=m, n,... or DEVIATION[c] or
        SIMPLE [c] or HELMERT or RHELMERT
SPECIFY hypothesis language
AMATRIX [matrix]
CMATRIX [matrix]
DMATRIX [matrix]
POST grpvariable
PAIRWISE
ERROR [matrix]
TEST
```

## Usage Considerations

**Types of data.** Normally, you analyze raw cases-by-variables data with the MANOVA module. You can, however, use a symmetric matrix data file (for example, a covariance matrix saved in a file from Correlations) as input. If you use a matrix as input, you must specify a value for Cases when estimating the model (under Model options in the



MANOVA Model tab) to specify the sample size of the data file that generated the matrix. The value in the dialog must be greater than 2.

SYSTAT uses the sample size to calculate degrees of freedom in hypothesis tests. SYSTAT also determines the type of matrix (SSCP, covariance, and so on) and adjusts appropriately. With a correlation matrix, the raw and standardized coefficients are the same; therefore, you cannot include a constant when using SSCP, covariance, or correlation matrices. Because these matrices are centered, the constant term has already been removed. If you give the sample size '2' you may get the residual degrees of freedom as zero.

**Print options.** The MANOVA module produces extended output if you set the output length to LONG.

For model estimation, the extended output adds the following: total sum of squares and product matrix, residual (or pooled within groups) sum of product matrix, residual (or pooled within groups) covariance matrix, and the residual (or pooled within groups) correlation matrix.

For hypothesis testing, the extended output adds **A**, **C**, and **D** matrices, the matrix of contrasts, and the inverse of the cross products of contrasts, hypothesis and error sum of product matrices, tests of residual roots, canonical correlations, and coefficients.

**Quick Graphs.** If no variables are categorical, MANOVA produces Quick Graphs of residuals versus predicted values.

**Saving files.** Several sets of the output can be saved to a file. The actual contents of the saved file depend on the analysis. Files may include the estimated regression coefficients, model variables, residuals, predicted values and diagnostic statistics.

**BY groups.** Each level of any BY variables yields a separate analysis.

**Case frequencies.** MANOVA uses the FREQUENCY variable, if present, to duplicate cases.

**Case weights.** MANOVA uses the values of any WEIGHT variables to weight each case.

## Examples

### Example 1 One-Way MANOVA

Here is an example from Jackson (2003), which deals with one-way classified data where samples were tested in three different laboratories using two different methods. In each laboratory two methods were used to test samples of size four. In one laboratory a sample of eight observations was tested. We can perform a multivariate analysis on this data to test for differences in laboratories. Here the dependent variables are *METHOD1* and *METHOD2*.

The input is:

```
USE LAB
PLENGTH SHORT
MANOVA
CATEGORY LAB / EFFECT
MODEL METHOD1 METHOD2 = CONSTANT + LAB
ESTIMATE
```

The output is:

#### Dependent Variable Means

METHOD1	METHOD2
10.275	10.083

#### Estimates of Effects $B = (X'X)^{-1}X'Y$

Factor	Level	METHOD1	METHOD2
CONSTANT		10.275	10.083
LAB	1	-0.275	0.267
LAB	2	-0.275	-0.083

#### Information Criteria

AIC	22.977
AIC (Corrected)	112.977
Schwarz's BIC	27.341

The above table displays the estimated effects for the fitted model; the means and information criteria table are also displayed above.

## Test of Hypothesis

Our interest is to simultaneously compare the three laboratories. We test the effect of LAB.

The input is:

HYPOTHESIS

EFFECT LAB

TEST

The output is:

Test for effect called: LAB

Null Hypothesis Contrast AB

	METHOD1	METHOD2
1	-0.275	0.267
2	-0.275	-0.083

Inverse Contrast  $A(X'X)^{-1}A'$

	1	2
1	0.167	
2	-0.083	0.167

Hypothesis Sum of Product Matrix  $H = B'A'(A(X'X)^{-1}A')^{-1}AB$

	METHOD1	METHOD2
METHOD1	1.815	
METHOD2	-0.605	0.447

Error Sum of Product Matrix  $G = E'E$

	METHOD1	METHOD2
METHOD1	2.728	
METHOD2	2.630	2.810

Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
METHOD1	1.815	2	0.908	2.995	0.101
Error	2.728	9	0.303		
METHOD2	0.447	2	0.223	0.715	0.515
Error	2.810	9	0.312		

Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.070	11.130	4, 16	0.000
Pillai Trace	0.972	4.252	4, 18	0.013
Hotelling-Lawley Trace	12.712	22.246	4, 14	0.000

THETA	S	M	N	p-value
0.927	2	-0.500	3.000	0.000

From the above table we get the four multivariate test statistics corresponding to the null hypothesis. In this example Wilks's  $\Lambda$  is 0.070 and its F approximation is 11.130. The corresponding *p-value* (less than 0.05) implies that there is sufficient evidence against the null hypothesis. All the remaining four tests reveal the same result. The exact test procedure and table value for Roy's greatest root are also displayed.

## Example 2

### Two-Way MANOVA

The data in the file *MANOVA* contains results of a hypothetical experiment on mice assigned randomly to one of three drugs. The weight loss in grams was observed for the first and second weeks of the experiment. The data were analyzed in Morrison (2004) with a two-way multivariate analysis of variance (a two-way MANOVA).

The input is:

```
USE MANOVA
MANOVA
CATEGORY SEX, DRUG / EFFECT
MODEL WEEK(1 .. 2) = CONSTANT + SEX + DRUG + SEX*DRUG
PLENGTH SHORT
ESTIMATE
```

The output is:

#### Dependent Variable Means

WEEK(1)	WEEK(2)
9.750	8.667

#### Estimates of Effects $B = (X'X)^{-1}X'Y$

Factor	Level	WEEK(1)	WEEK(2)
CONSTANT		9.750	8.667
SEX	1	0.167	0.167
DRUG	1	-2.750	-1.417
DRUG	2	-2.250	-0.167
SEX*DRUG	1*1	-0.667	-1.167
SEX*DRUG	1*2	-0.417	-0.417

#### Information Criteria

AIC	217.701
AIC (Corrected)	277.701
Schwarz's BIC	235.372



Notice that each column of the B matrix is now assigned to a separate dependent variable. It is as if we had done two runs of an ANOVA. The numbers in the matrix are the analysis of variance effects estimates.

### *Test of Hypotheses*

You can test the following three hypotheses. The extended output for the second hypothesis is used to illustrate the detailed output.

The input is:

```
HYPOTHESIS
EFFECT SEX
PLENGTH LONG
TEST
```

```
HYPOTHESIS
EFFECT DRUG
PLENGTH SHORT
TEST
```

```
HYPOTHESIS
EFFECT SEX*DRUG
TEST
```

The output is:

Test for effect called: SEX

Null Hypothesis Contrast AB

WEEK (1)	WEEK (2)
0.167	0.167

Inverse Contrast  $A(X'X)^{-1}A'$

0.042

Hypothesis Sum of Product Matrix  $H = B'A'(A(X'X)^{-1}A')^{-1}AB$

	WEEK (1)	WEEK (2)
WEEK (1)	0.667	
WEEK (2)	0.667	0.667

Error Sum of Product Matrix  $G = E'E$ 

	WEEK(1)	WEEK(2)
WEEK(1)	94.500	
WEEK(2)	76.500	114.000

## Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
WEEK(1)	0.667	1	0.667	0.127	0.726
Error	94.500	18	5.250		
WEEK(2)	0.667	1	0.667	0.105	0.749
Error	114.000	18	6.333		

## Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.993	0.064	2, 17	0.938
Pillai Trace	0.007	0.064	2, 17	0.938
Hotelling-Lawley Trace	0.008	0.064	2, 17	0.938

## Test of Residual Roots

Roots	Chi-square	df
1 through 1	0.157	2

## Canonical Correlations

0.086

Dependent Variable Canonical Coefficients Standardized  
by Conditional (within Groups) Standard Deviations

WEEK(1)	0.698
WEEK(2)	0.368

Canonical Loadings (Correlations between Conditional  
Dependent Variables and Dependent Canonical Factors)

WEEK(1)	0.969
WEEK(2)	0.882

## Test for effect called: DRUG

## Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
WEEK(1)	301.000	2	150.500	28.667	0.000
Error	94.500	18	5.250		
WEEK(2)	36.333	2	18.167	2.868	0.083
Error	114.000	18	6.333		

## Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.169	12.199	4, 34	0.000
Pillai Trace	0.880	7.077	4, 36	0.000
Hotelling-Lawley Trace	4.640	18.558	4, 32	0.000

THETA	S	M	N	p-value
0.821	2	-0.500	7.500	0.000

Test for effect called: SEX\*DRUG

## Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
WEEK(1)	14.333	2	7.167	1.365	0.281
Error	94.500	18	5.250		
WEEK(2)	32.333	2	16.167	2.553	0.106
Error	114.000	18	6.333		

## Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.774	1.159	4, 34	0.346
Pillai Trace	0.227	1.152	4, 36	0.348
Hotelling-Lawley Trace	0.290	1.159	4, 32	0.347

THETA	S	M	N	p-value
0.221	2	-0.500	7.500	0.295

Matrix formulae (that are sometimes long) make explicit the hypothesis being tested. For MANOVA, hypotheses are tested with sum of squares and cross-products matrices. Before printing the multivariate tests, however, SYSTAT prints the univariate tests. Each of these *F-ratios* is constructed in the same way as in ANOVA model. The sum of squares for the hypothesis and error are taken from the diagonals of the respective sum of squares and product matrices. The univariate F test for the *WEEK(1) DRUG* effect, for example, is computed from  $301.0 / 2$  over  $94.5 / 18$ , or hypothesis mean square divided by error mean square.

The next statistics printed are for the multivariate hypothesis. Wilks's lambda (likelihood-ratio criterion) varies between 0 and 1. Schatzoff (1966) has tables for its percentage points. The following F-ratio is Rao's approximate (sometimes exact) F statistic corresponding to the likelihood-ratio criterion (see Rao, 1973). Pillai's trace and its F approximation are taken from Pillai (1960). The Hotelling-Lawley trace and its F approximation are documented in Morrison (2004). The last statistic is the largest root criterion for Roy's union-intersection test (see Morrison, 2004). Charts of the percentage points of this statistic, found in Morrison and other multivariate texts, are taken from Heck (1960).

The probability value printed for THETA is not an approximation. It is what you find in the charts. In the first hypothesis, all the multivariate statistics have the same value for the F approximation because the approximation is exact when there are only two groups (see Hotelling's  $T^2$  in Morrison, 2004). In these cases, THETA is not printed because it has the same probability value as the F-ratio.



### ***Bartlett's Residual Root (Eigenvalue) Test***

The chi-square statistics follow Bartlett (1947). The probability value for the first chi-square statistic should correspond to that for the approximate multivariate *F-ratio* in large samples. In small samples, they might be discrepant, in which case you should generally trust the *F-ratio* more. The subsequent chi-square statistics are recomputed, leaving out the first and later roots until the last root is tested. These are sequential tests and should be treated with caution, but they can be used to decide how many dimensions (roots and canonical correlations) are significant. The number of significant roots corresponds to the number of significant *p-values* in this ordered list.

### ***Canonical Coefficients***

Dimensions with insignificant chi-square statistics in the prior tests should be ignored in general. Corresponding to each canonical correlation is a canonical variate, whose coefficients have been standardized by the within-groups standard deviations (the default). Standardization by the sample standard deviation is generally used for canonical correlation analysis or multivariate regression when groups are not present to introduce covariation among variates. You can standardize these variates by the total (sample) standard deviations with:

STANDARDIZE TOTAL

inserted prior to TEST. Continue with the other test specifications described earlier.

Finally, the canonical loadings are printed. These are correlations and, thus, provide information different from the canonical coefficients. In particular, you can identify suppressor variables in the multivariate system by looking for differences in sign between the coefficients and the loadings (which is the case with these data). See Bock (1975) and Wilkinson (1975, 1977) for an interpretation of these variates.

Since the equality of means for the effect called *DRUG* is rejected, our next concern will be to find the pair of drugs which differ more significantly. You can perform all pairs comparisons by choosing from the menu:

Analyze

MANOVA

Between-Group Testing...

In the dialog box under Specified effect select **EFFECT = DRUG** and check 'Perform all pairs Comparison'. Specify the Error term as **Model SSCP matrix**.



The input is:

```

HYPOTHESIS
POST DRUG
TEST

```

The output is:

```

All-pairs Comparison

```

SEX(i)	SEX(j)	Hotelling's T-square	p-value
1.000	2.000	0.135	0.938

### Example 3 Multivariate Nested Design

We consider an example of a nested design (Timm, 2002) in which teachers are nested within classes. The design for this analysis would be a fixed effects nested design with more than one response variable.

The input is:

```

USE TEACHER
MANOVA
CATEGORY CLASSES$ TEACHERS$ / EFFECT
MODEL READRATE READCOMP = CONSTANT + CLASSES$ +,
TEACHERS$(CLASSES$)
ESTIMATE

```

The output is:

Test for effect called: CLASSES\$

Null Hypothesis Contrast AB

READRATE	READCOMP
-0.550	-2.383

Inverse Contrast  $A(X'X)^{-1}A'$

0.042

Hypothesis Sum of Product Matrix  $H = B'A'(A(X'X)^{-1}A')^{-1}AB$

	READRATE	READCOMP
READRATE	7.260	
READCOMP	31.460	136.327

Error Sum of Product Matrix  $G = E'E$ 

	READRATE	READCOMPRE
READRATE	42.800	
READCOMPRE	20.800	42.000

## Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
READRATE	7.260	1	7.260	3.393	0.080
Error	42.800	20	2.140		
READCOMPRE	136.327	1	136.327	64.917	0.000
Error	42.000	20	2.100		

## Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.220	33.623	2, 19	0.000
Pillai Trace	0.780	33.623	2, 19	0.000
Hottelling-Lawley Trace	3.539	33.623	2, 19	0.000

Test for effect called: TEACHERS\$(CLASSES\$)

## Null Hypothesis Contrast AB

	READRATE	READCOMPRE
1	-0.500	-0.500
2	0.200	0.133
3	2.600	3.733

Inverse Contrast  $A(X'X)^{-1}A'$ 

	1	2	3
1	0.100		
2	0.000	0.133	
3	0.000	-0.067	0.133

Hypothesis Sum of Product Matrix  $H = B'A'(A(X'X)^{-1}A')^{-1}AB$ 

	READRATE	READCOMPRE
READRATE	75.700	
READCOMPRE	105.300	147.033

Error Sum of Product Matrix  $G = E'E$ 

	READRATE	READCOMPRE
READRATE	42.800	
READCOMPRE	20.800	42.000

## Univariate F Tests

Source	Type III SS	df	Mean Squares	F-ratio	p-value
READRATE	75.700	3	25.233	11.791	0.000
Error	42.800	20	2.140		
READCOMPRE	147.033	3	49.011	23.339	0.000
Error	42.000	20	2.100		

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.210	7.487	6, 38	0.000
Pillai Trace	0.796	4.412	6, 40	0.002
Hotelling-Lawley Trace	3.730	11.191	6, 36	0.000

THETA	S	M	N	p-value
0.788	2	0.000	8.500	0.000

#### Example 4

##### Repeated Measures Analysis in the Presence of Subject-Specific Covariates

The input is:

```
USE PHYSICAL
MANOVA
CATEGORY GROUP
MODEL Y1 Y2 Y3 Y4 = CONSTANT+GROUP+X1+X2+
X1*GROUP+X2*GROUP/REPEAT =4(1 2 3 4),
NAMES='Time'
PLENGTH SHORT
ESTIMATE
```

The output is:

Multivariate Repeated Measures Analysis

Test of: Time

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.926	3	11	0.293	0.830
Pillai Trace	0.074	3	11	0.293	0.830
Hotelling-Lawley Trace	0.080	3	11	0.293	0.830

Test of: Time\*GROUP

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.676	9	26	0.523	0.845
Pillai Trace	0.358	9	39	0.587	0.800
Hotelling-Lawley Trace	0.431	9	29	0.463	0.887

THETA	S	M	N	p-value
0.210	3	-0.500	4.500	0.698

Test of: Time\*X1

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.928	3	11	0.283	0.837
Pillai Trace	0.072	3	11	0.283	0.837
Hotelling-Lawley Trace	0.077	3	11	0.283	0.837

Test of: Time\*X2

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.966	3	11	0.129	0.941
Pillai Trace	0.034	3	11	0.129	0.941
Hotelling-Lawley Trace	0.035	3	11	0.129	0.941

Test of: Time\*GROUP\*X1

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.816	9	26	0.260	0.980
Pillai Trace	0.192	9	39	0.295	0.972
Hotelling-Lawley Trace	0.216	9	29	0.231	0.987

THETA	S	M	N	p-value
0.135	3	-0.500	4.500	0.531

Test of: Time\*GROUP\*X2

Statistic	Value	Hypothesis df	Error df	F-ratio	p-value
Wilks's Lambda	0.640	9	26	0.601	0.785
Pillai Trace	0.377	9	39	0.622	0.771
Hotelling-Lawley Trace	0.535	9	29	0.575	0.807

THETA	S	M	N	p-value
0.325	3	-0.500	4.500	0.624

None of the multivariate tests for Time\*X1, Time\*X2, Time\*X1\*GROUP, Time\*X2\*GROUP appears to be significant.



### Example 5

#### Within-Group Testing

In a clinical trial experiment (Crowder and Hand, 1990), two drug treatments, both in tablet form, were compared using five volunteer subjects in a pilot trial. There were two phases: in the first phase Drug A was used, and in the second phase Drug B was used. In each phase, the blood samples were taken at times 1, 2, 3, and 6 hours after medication and the resulting antibiotic serum levels were reported. We can perform a repeated measures analysis on this data set. We can fit a general linear model as follows,

The input is:

```
USE SERUM
MANOVA
CATEGORY DRUG$ / EFFECT
MODEL TIME1 TIME2 TIME3 TIME6 = CONSTANT + DRUG$
ESTIMATE
```

The output is:

#### Multivariate Test Statistics

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.797	0.319	4, 5	0.854
Pillai Trace	0.203	0.319	4, 5	0.854
Hotelling-Lawley Trace	0.255	0.319	4, 5	0.854

Here Wilks's  $\Lambda = 0.797$  and its corresponding  $p\text{-value} = 0.854$ ; this implies that there is no significant evidence to reject the null hypothesis. We can conclude that there is no difference between phases A and B.

Another question of interest is: Within each phase, are the antibiotic serum levels taken at four different times equal? Let us consider phase A.

The input is:

```
HYPOTHESIS
EFFECT DRUG$
AMATRIX [ 1 1 ]
CMATRIX [ -1 1 0 0; -1 0 1 0; -1 0 0 1 ]
TEST
```

The output is:

**Multivariate Test Statistics**

Statistic	Value	F-ratio	df	p-value
Wilks's Lambda	0.088	20.732	3, 6	0.001
Pillai Trace	0.912	20.732	3, 6	0.001
Hotelling-Lawley Trace	10.366	20.732	3, 6	0.001

Wilks's  $\Lambda=0.088$  and its  $p$ -value is 0.001 showing that there is significant evidence to reject the null hypothesis. Hence we may conclude that within phase A the antibiotic serum levels taken at four different times are not the same. Similarly you can perform the above-mentioned test for phase B also.

The input is:

```

HYPOTHESIS
EFFECT DRUG$
AMATRIX [ 1 -1 ]
CMATRIX [ -1 1 0 0;-1 0 1 0;-1 0 0 1 ]
TEST

```

Instead of using commands you can use the dialog box for within group testing. If you use this, there is no need to specify the A matrix. The above-mentioned hypothesis for phase A can be done through the dialog box by checking "Equality of means". In such a case you do not have to specify the matrices A and C.

### Example 6

#### AIC and Schwarz's BIC

The data set in *ROHWER* consists of the performance of 32 kindergartens in three standardized tests: Peabody Picture Vocabulary Test (*PPVT*), Raven Progressive Matrices Test (*RPMT*), and Student Achievement Test (*SAT*). The independent variables are: Named (*N*), Still (*S*), Named Still (*NS*), Named Action (*NA*), and Sentence Still (*SS*).

This data set illustrates how information criteria can be employed as a tool for model selection.

In this example, analysis is performed by fitting all possible sub-models, and the corresponding information criteria are obtained. All possible sub-models are fitted by

executing the commands in the command file *MULTIVARIATE REGRESSION.SYC*. The command script for fitting a candidate sub-model is as follows:

```
MANOVA
USE ROHWER
MODEL PPVT RPMT SAT = CONSTANT + NA
ESTIMATE
MODEL PPVT RPMT SAT = CONSTANT + S + NS + NA
ESTIMATE
```

The following table presents the models with low information criteria values.

Model number	Model terms	AIC	AIC (corrected)	Schwarz's BIC
1	CONSTANT+NA	723.955	740.376	720.750
2	CONSTANT+S+NS+NA	718.159	770.775	723.7483

Model 1 corresponds to smaller AIC (corrected) and Schwarz's BIC among all possible candidate sub-models. Model 2 is the model corresponding to smaller AIC value among all possible sub-models. The AIC value for the model with *CONSTANT*, *N*, *S*, *NS* and *NA* as independents is close to the AIC value for the Model 1.

From the analysis, it appears that Model 1 is a better approximation of the true model among all possible sub-models.

## References

- Bartlett, M.S. (1947). Multivariate analysis. *Journal of the Royal Statistical Society*, Series B, 9, 176-197.
- Bock, R.D. (1975). *Multivariate statistical methods in behavioral research*. New York: McGraw-Hill.
- Crowder, M. J. and Hand, D.J. (1990). *Analysis of repeated measures*. London: Chapman & Hall.
- Heck, D.L. (1960). Charts of some upper percentage points of the distribution of the largest characteristic root. *Annals of Mathematical Statistics*, 31, 625-642.
- Jackson, J.E. (2003). *A user's guide to principal components*. New York: Wiley-Interscience.
- Johnson, R.A. and Wichern, D.W. (2002). *Applied multivariate statistical analysis*, 5th ed. Englewood Cliffs, N.J.: Prentice Hall.
- Morrison, D. F. (2004). *Multivariate statistical methods*, 4th ed. Pacific Grove, CA:



Duxbury Press.

Pillai, K. C. S. (1960). *Statistical table for tests of multivariate hypotheses*. Manila: The Statistical Center, University of Philippines.

Rao, C.R. (1973). *Linear statistical inference and its applications*, 2nd ed. New York: John Wiley & Sons.

Rencher, A.C. (2002). *Methods of multivariate analysis*, 2nd ed. New York: John Wiley & Sons.

Schatzoff, M. (1966). Exact distributions of Wilks's likelihood ratio criterion. *Biometrika*, 53, 347-358.

Timm, N.H. (2002). *Applied multivariate analysis*. New York: Springer-Verlag.

Wilkinson, L. (1975). Response variable hypotheses in the multivariate analysis of variance. *Psychological Bulletin*, 82, 408-412.

Wilkinson, L. (1977). Confirmatory rotation of MANOVA canonical variates. *Multivariate Behavioral Research*, 12, 487-494.



# Nonlinear Models

Laszlo Engelman

Nonlinear modeling estimates parameters for a variety of nonlinear models using a Gauss-Newton (SYSTAT computes exact derivatives), Quasi-Newton, or Simplex algorithm. In addition, you can specify a loss function other than least-squares, so maximum likelihood estimates can be computed. You can set lower and upper limits on individual parameters. When the parameters are highly intercorrelated, and there is concern about overfitting, you can fix the value of one or more parameters, and Nonlinear Model will test the result against the full model. If the estimates have trouble converging, or if they converge to a local minimum, Marquarding is available.

For assessing the certainty of the parameter estimates, Nonlinear Model offers Wald confidence regions and Cook and Weisberg (1990) confidence curves. The latter are useful when it is unreasonable to assume that the estimates follow a normal distribution. You can also save values of the loss function for plotting contours in a bivariate display of the parameter space. This allows you to study the combinations of parameter estimates with approximately the same loss function values.

When your response contains outliers, you may want to downweight their residuals using one of Nonlinear Model's robust  $\psi$  functions: median, Huber, trim, Hampel,  $t$ , Bisquare, Ramsay, Andrews, Tukey, or the  $p^{\text{th}}$  power of the absolute value of the residuals.

You can specify functions of parameters (like LD50 for a logistic model). SYSTAT evaluates the function at each iteration, and prints the standard error and the Wald interval for the estimate after the last iteration.

Resampling procedures are available in this feature.

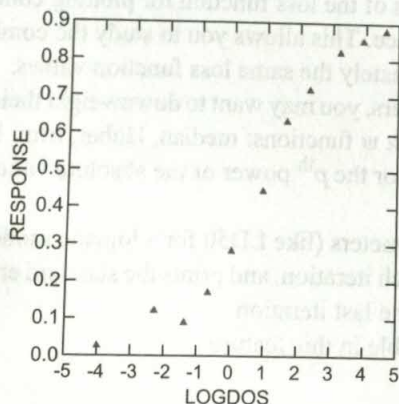
## Statistical Background

The following data are from a toxicity study for a drug designed to combat tumors. The table shows the proportion of laboratory rats dying (*Response*) at each dose level (*Dose*) of the drug. Clinical studies usually scale dose in natural logarithm units, which are listed in the center column (*Log Dose*). We arbitrarily set the *Log Dose* to -4 for zero *Dose* to be able to plot and fit a linear model.

Dose	Log Dose	Response
0.00	-4.000	0.026
0.10	-2.303	0.120
0.25	-1.386	0.088
0.50	-0.693	0.169
1.00	0.000	0.281
2.50	0.916	0.443
5.00	1.609	0.632
10.00	2.303	0.718
25.00	3.219	0.820
50.00	3.912	0.852
100.00	4.605	0.879

## Modeling the Dose-Response Function

The plot of *Response* against *LOGDOS* (*Log Dose*) is clearly curvilinear.



The S-shaped function suggests that we could use a linear model with linear, quadratic, and cubic terms (that is, a polynomial function) to fit a curved line to the data. Here are the results:

```
Dependent Variable      RESPONSE
N                        11
Multiple R               0.993
Squared Multiple R       0.986
Adjusted Squared Multiple R 0.980
Standard Error of Estimate 0.047
```

Regression Coefficients B =  $(X'X)^{-1}X'Y$

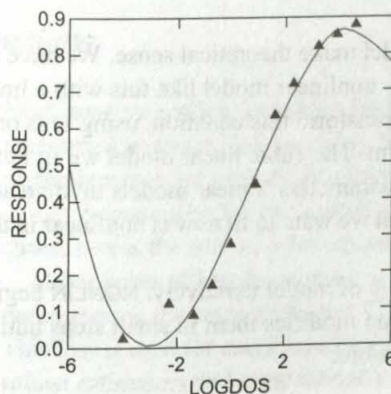
Effect	Coefficient	Standard Error	Std. Coefficient	Tolerance
CONSTANT	0.314	0.021	0.000	.
LOGDOS	0.166	0.013	1.344	0.168
LOGDOS*LOGDOS	0.009	0.002	0.202	0.771
LOGDOS*LOGDOS*LOGDOS	-0.004	0.001	-0.492	0.152

Regression Coefficients B =  $(X'X)^{-1}X'Y$  (contd...)

Effect	t	p-value
CONSTANT	15.241	0.000
LOGDOS	12.418	0.000
LOGDOS*LOGDOS	3.995	0.005
LOGDOS*LOGDOS*LOGDOS	-4.322	0.003

Notice that all the coefficients are highly significant and the overall fit is excellent ( $R^2 = 0.986$ ). Even the tolerances are relatively large, so we need not worry about collinearity. The residual plots for this function are reasonably well behaved. There is no significant autocorrelation in the residuals.

The following figure shows the observed data and the fitted curve.





How do the researchers interpret this plot? First of all, the curve is consistent with the printed output; it fits extremely well in the range of the data. Putting the fitted curve into ordinary language, we can say that fewer animals die at lower dosages and more at higher. At the extremes, however, more animals die with extremely low dosages and fewer animals die at extremely high dosages.

This is nonsense. While it is possible to imagine some drugs (arsenic, for example) for which dose-response functions are nonmonotonic, the model we fit makes no sense for a clinical drug of this sort. Second, the cubic function we fit extrapolates beyond the 0–1 response interval. It implies that there is something beyond dying and something less than living. Third, the parameters of the model we fit have no theoretical interpretation.

Clinical researchers usually prefer to fit quantal response data like these with a bounded monotonic response function of the following form:

$$\text{proportion dying} = \alpha + \frac{1 - \alpha}{1 + e^{[\beta - \gamma \log(\text{dose})]}}$$

where  $\alpha$  is the background response, or rate of dying,  $\beta$  is a location parameter for the curve, and  $\gamma$  is a slope parameter for the curve.

Estimating a quantity called LD50 is the usual purpose of this type of study. LD50 is the dose at which 50 percent of the animals are expected to die. LD50 is:

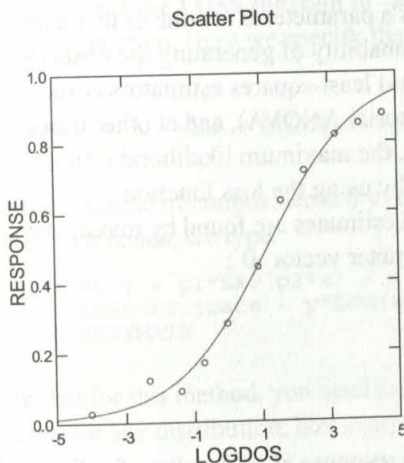
$$e^{\beta/\gamma} (1 - 2\alpha)^{1/\gamma}$$

Notice how the parameters of this model make theoretical sense. We have a problem, however. We cannot fit an intrinsically nonlinear model like this with a linear regression program. We cannot even transform this equation, using logs or other mathematical operators, to a linear form. The cubic linear model we fit before was nonlinear in the data but linear in the parameters. Linear models involve additive combinations of parameters. The model we want to fit now is nonlinear in the data and nonlinear in the parameters.

We need a program that fits this type of model iteratively. **NONLIN** begins with initial estimates of parameter values and modifies them in small steps until the fit of the curve to the data is as close as possible.



Here is the result:



Notice how the curve tapers at the ends so that it is bounded by 0 and 1 on the *Response* scale. This behavior fits our theoretical ideas about the effect of this drug. The value for LD50 is 3.262, which is in raw dose units.

Interestingly, this model does not fit significantly better than the cubic polynomial. Both have comparable sum of squared residuals. True, the cubic model has four parameters and we have used only three. Nevertheless, this example should convince you that blind searching for models that produce good fits is not good science. It is even possible that a model with a poorer fit can be the true model generating data and one with a better fit can be bogus.

## Loss Functions

Nonlinear estimation includes a broad variety of statistical procedures. We have performed nonlinear least-squares, which is analogous to ordinary least-squares. Both methods minimize squared deviations of the dependent variable data values from values estimated by the function at the same independent variable data points. In these cases, loss is the sum of least-squares.

Other types of loss functions can be defined which produce different estimates of parameters in the same functions. The most widely used loss is negative log likelihood. This loss is used for maximum likelihood estimation. Other loss functions are used for robust estimators and nonparametric procedures.

### Maximum Likelihood

A maximum likelihood estimate of a parameter is a value of that parameter in a given distribution that has the highest probability of generating the observed sample data. Sometimes maximum likelihood and least-squares estimators coincide (as in fixed effects, fully crossed, balanced factorial ANOVA), and at other times they diverge. In our quantal response data example, the maximum likelihood estimates are different. They can be computed in NONLIN by using the loss function.

In general, maximum likelihood estimates are found by maximizing the likelihood function  $L$  with respect to the parameter vector  $\theta$  :

$$L = \prod_{i=1}^n d(x_i, \theta)$$

where  $d(x_i, \theta)$  is the density of the response at each value of  $x$ . Equivalently, the negative of the log of the likelihood function can be minimized:

$$-\log L = -\sum_{i=1}^n \ln(d(x_i, \theta))$$

Here we outline four methods for computing maximum likelihood estimates in NONLIN. To define them, we use a specific model and a specific density. The model is the sum of two exponentials:

$$\hat{y} = p_1 e^{p_2 x} + p_3 e^{p_4 x}$$

and the distribution of  $y$  at each  $x$  is Poisson:

$$d(x_i, \lambda) = \frac{e^{-\lambda} \lambda^{y_i}}{y_i!}$$

In our definitions, we also use the log of the density:

$$\ln d = -\lambda + y \ln \lambda - \text{LGM}(y + 1)$$

where LGM is the log gamma function for computing  $y!$ .

**Method 1. Set the LOSS function to  $-\ln(\text{density})$ .** In NONLIN, you can specify your own loss function. Here we specify the negative of the log of the density function:

$$\text{LOSS} = \lambda - y \ln \lambda + \text{LGM}(y + 1)$$

For the estimate of lambda, we use  $\hat{y}$ , or **estimate**, as it is known to Nonlinear Model. Using commands, we type:

```
MODEL Y = p1*EXP(p2*x) + p3*EXP(p4*x)
LOSS estimate - y*LOG(estimate) + LGM(y+1)
ESTIMATE
```

Note that for this method, you need to specify only the loss function. This method can be used for any distribution; however, the estimated standard errors may not be correct.

**Method 2. Iteratively reweighted least-squares.** This method is appropriate for distributions belonging to the exponential family (for example, normal, binomial, multinomial, Poisson, and gamma). It provides meaningful standard errors for the parameter estimates and useful residuals. For this method, you define a case weight that is recomputed at each iteration:

$$\text{weight} = \frac{1}{\text{variance}(y_i)}$$

For our Poisson distribution, the mean and variance are equal, so lambda is the variance, and our estimate of the variance is *estimate*. Thus, the weight is:

$$\text{weight} = \frac{1}{\text{estimate}}$$

Here's how to specify this method using NONLIN commands:

```
LET wt=1
WEIGHT wt
MODEL y = p1*EXP(p2*x) + p3*EXP(p4*x)
RESET wt = 1 / estimate
ESTIMATE / SCALE
```

The standard deviation of the resulting estimates are the usual information theory standard errors.



**Method 3. Estimate  $\ln(\text{density})$  and reset the predicted value to  $y + 1$ .** For this method, the data may follow any distribution and the standard errors are correct, but the method does not yield correct residuals. You define a dummy outcome variable and estimate the log of the density, and then reset the outcome variable to  $\hat{y} + 1$  at each iteration. For our example, use the commands:

```
LET dummy = 0
MODEL dummy = -p1*EXP(p2*x) - p3*EXP(p4*x),
               + y*LOG(p1*EXP(p2*x) + p3*EXP(p4*x)),
               -LGM(y + 1)
RESET dummy = estimate + 1
ESTIMATE / SCALE
```

**Method 4. Set the predicted value to zero and define the function as the square root of the negative log density.** This method is a variation of method 1, so it is appropriate for data from any distribution and provides estimates of the parameters only. Here we trick NONLIN by setting  $y=0$  for all cases:

$f = \sqrt{-\ln d(x, \theta)}$ , so  $\Sigma(y - f)^2$  becomes

$$\Sigma(0 - \sqrt{-\ln d(x, \theta)})^2 = \Sigma -\ln d(x, \theta)$$

For our example, use the commands:

```
LET dummy = 0
MODEL dummy = SQR(p1*EXP(p2*x) + p3*EXP(p4*x)),
               - y*LOG(p1*EXP(p2*x) + p3*EXP(p4*x)),
               + LGM(y + 1)
ESTIMATE
```

### ***Least Absolute Deviations***

As an example of other types of loss functions, consider minimizing least absolute values of deviations of the dependent variable data values from values estimated by the function at the same independent variable data points. This procedure produces estimates which, on the average, are influenced less by outliers than the least-squares estimates. This is because squaring a large value increases its impact. While there are more sophisticated robust procedures, least absolute values estimates are easy to compute in NONLIN and fun to compare with least-squares estimates.



## ***Model Estimation***

SYSTAT provides three algorithms for estimating your model: Gauss-Newton, Quasi-Newton, and Simplex. The Gauss-Newton method with its exact derivatives produces more accurate estimates of the asymptotic standard errors and covariances and can converge in fewer iterations and more quickly than the other two algorithms.

Both GN and the Quasi-Newton method do not work if the derivatives are undefined in the region in which you are seeking minimum values. Specifically, the first and second derivatives must exist at all points for which the algorithm computes values. However, the algorithms cannot identify situations where the derivatives do not exist. Also, Quasi-Newton cannot detect when derivatives fluctuate rapidly—thus, Gauss-Newton can be more accurate.

The Simplex algorithm does not have this requirement. It calculates a value for your loss function at some point, looks to see if this value is less than values elsewhere, and steps to a new point to try again. When the steps become small, iterations stop.

GN is the fastest method. Simplex is generally slower than the others, particularly for least-squares, because Simplex cannot make use of the information in the derivatives to find how far to move its estimates at each step.

## ***How Nonlinear Modeling Works***

The estimation works as follows: the starting values of the parameters are selected by the program or by you. The model (if stated) is then evaluated for the first case in double precision. The result of this function is called the estimate. Next, the loss function is evaluated for the first case, using the estimate from the model. If you did not include a loss function, then loss is computed by squaring the residual for the first case.

This procedure is repeated for all cases in the file and the loss is summed over cases. The summed loss is then minimized using the Gauss-Newton, Quasi-Newton, or Simplex algorithms. Iterations continue until both convergence criteria are met or the maximum number of iterations is reached.

## ***Problems***

You may encounter numerous pitfalls (for example, dependencies, discontinuities, local minima, and so on). Nonlinear Model offers several possibilities to overcome these pitfalls, but, in some instances, even your best efforts may be futile.

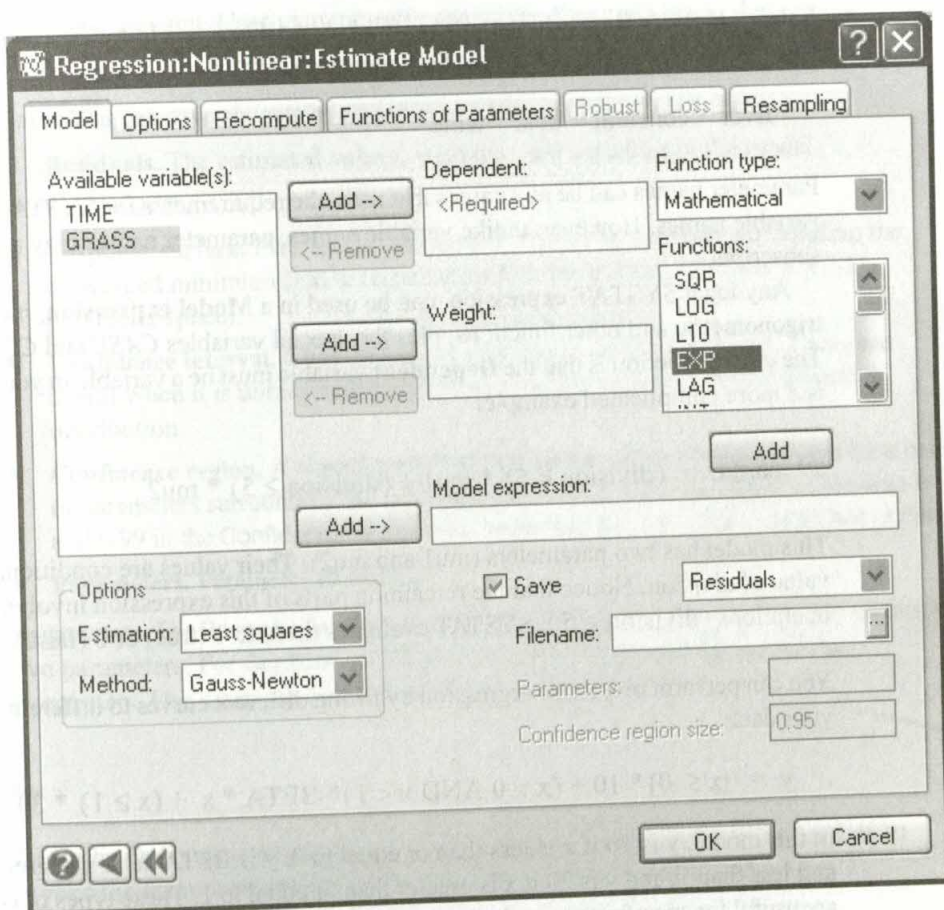
- Find reasonable starting values by considering approximately what the values should be. Try plotting the data. For example in the contouring example, you could let  $DAYS \rightarrow \infty$  and estimate  $\theta_1$  to be approximately 20.
- Try Marquardingt.
- Use several different starting values for each method before you feel comfortable with the final estimates. This can help you expose local minima. The Simplex method is the most robust against local minima. There is a trade-off, however, because it is considerably slower.
- Try switching back and forth between Gauss-Newton, Quasi-Newton, and Simplex without changing the starting values. That way, one may help you out of a convergence or local minimum problem.
- If you get illegal function values for starting values, try some other estimates. For some functions with many parameters, you may need high quality starting values to even get an estimable function!
- Never trust the output of an iterative nonlinear estimation procedure until you have plotted estimates against predictors and you have tried several different starting values. SYSTAT is designed so that you can quickly save estimates, residuals, and model variables and plot them. All of the examples in this chapter were tested this way. Although most began with default starting values for the parameters, they were checked with other starting values.

## ***Nonlinear Models in SYSTAT***

### ***Nonlinear Regression: Estimate Model***

To open the Nonlinear Regression: Estimate Model dialog box, from the menus choose:

Analyze  
Regression  
Nonlinear  
Estimate Model...



**Model specification.** Specify a general algebraic model to be estimated. Terms that are not variables are assumed to be parameters. If you want to use a function in the model, choose a Function type from the drop-down list, select the function in the functions list, and click Add.

Nonlinear modeling uses models resembling those for General Linear Models (GLM). There is one critical difference, however. The Nonlinear Model statement is a literal algebraic expression of variables and parameters. Choose any name you want for these parameters. Any names you specify that are not variable names in your file are assumed to be parameter names. Suppose you specify the following model for the *USSTATES* data:

$$\text{liver} = b0 + b1 * \text{wine}$$



Select *LIVER* as **Dependent** and specify  $b_0 + b_1 * WINE$  as **Model expression**.

Since  $b_0$  and  $b_1$  are not variables (they are parameters), the following model is the same:

$$\text{liver} = \text{constant} + \text{beta} * \text{wine}$$

Parameter names can be any names that meet the requirements for SYSTAT's numeric variable names. However, unlike variable names, parameter names may not have subscripts.

Any legal SYSTAT expression can be used in a **Model expression**, including trigonometric and other functions, plus the special variables *CASE* and *COMPLETE*. The only restriction is that the **Dependent** variable must be a variable in your file. Here is a more complicated example:

$$\text{cardio} = (\text{division} < 5) * \mu_1 + (\text{division} \geq 5) * \mu_2$$

This model has two parameters ( $\mu_1$  and  $\mu_2$ ). Their values are conditional on the value of division. Notice that the remaining parts of this expression involve relational operations ( $\text{division} \geq 5$ ). SYSTAT evaluates these to 1 (true) or 0 (false).

You can perform piecewise regression by fitting different curves to different subsets of your data:

$$y = (x \leq 0) * 10 + (x > 0 \text{ AND } x < 1) * \text{BETA} * x + (x \geq 1) * 20$$

In this model,  $y$  is 10 if  $x$  is less than or equal to 0,  $y$  is  $\text{BETA} * x$  if  $x$  is greater than 0 and less than 1, and  $y$  is 20 if  $x$  is greater than or equal to 1. These types of constraints are useful for specifying bounded probability functions such as the cumulative uniform distribution;

**Weight.** Selects the variable as a weight variable, which is to be used for estimating parameters by Iteratively Reweighted Least-Squares.

**Estimation.** You can specify a loss function other than least-squares. From the drop-down list, select Loss function to perform loss analysis. When your response contains outliers, you may want to downweight their residuals using a robust  $\psi$  function by selecting Robust.

**Method.** Three model estimation methods are available.

- **Gauss-Newton.** Computes exact derivatives.



- **Quasi-Newton.** Uses numeric estimates of the first and second derivatives.
- **Simplex.** Uses a direct search procedure.

**Save.** You can save six sets of statistics to a file.

- **Residuals.** The estimated values, residuals, and variables in the model.
- **Residuals/Data.** All of the above.
- **Response surface.** Five levels of contours of the loss function surrounding the converged minimum (like a response surface for the loss function in a 2-D parameter space).
- **Confidence interval.** Cook-Weisberg graphical confidence curves. These are useful when it is unreasonable to assume that the estimates follow a normal distribution.
- **Confidence region.** A closed curve that defines the  $n\%$  confidence region for a pair of parameters surrounding the converged minimum. Type a number,  $n$ , between 0 and 0.99 in the Confidence region field to specify the size of the confidence region.
- **Parameters.** Parameter estimates.

**Parameters.** For Response surface and Confidence region, you must specify names of two parameters. For Confidence interval, you must specify the names of the parameters. Use a comma between each parameter name.

## Options

Click the Options tab in the Nonlinear Regression: Estimate Model dialog box to invoke the estimation options.

**Regression: Nonlinear: Estimate Model**

Model Options Recompute Functions of Parameters Robust Loss Resampling

Starting values:

Minimum:

Maximum:

Iterations:

Step-halvings:

Tolerance:

Loss convergence:

Parameter convergence:

Fix:

☒ Use Marquardt

☐ Mean square error scale

Navigation buttons: [Back] [Forward] [Help]

OK Cancel

SYSTAT offers several options for controlling model computation.

**Starting values.** Starting values for model parameters. Specify values for each parameter in the order the parameters appear in your model (or loss statement if no model is specified). Separate the values with commas or blanks. You can specify starting values for some of the parameters and leave blanks for others.

SYSTAT chooses starting values if you do not. Specify starting values that give the general shape of the function you expect as a result. For example, if you expect that the function is a negative exponential function, then specify initial values that yield a negative exponential function. Also, make sure that the starting values are in a reasonable range. For example, if the function contains  $\text{EXP}(P \cdot \text{TIME})$  and TIME ranges from 10,000 to 20,000, then the initial value of P should be around  $1/10,000$ . If you

specified an initial value such as 0.1, the function would have extremely large values, such as  $e^{1000}$ .

**Minimum.** Lower limits for the parameters, one number per parameter.

**Maximum.** Upper limits for the parameters, one number per parameter.

**Iterations.** Maximum number of iterations for fitting your model. Default value is 25.

**Step-halvings.** Maximum number of step halvings. If the loss increases between two iterations, Nonlinear Model halves the increment size, computes the loss at the midpoint, and compares it to the residual sum of squares at the previous iteration. This process continues until the residual sum of squares is less than that at the previous iteration or until the maximum number of halvings is reached.

**Tolerance.** A check for near singularity. SYSTAT cannot invert the matrix of sums of cross-products of the derivatives with respect to the parameters if the matrix is singular. Use Tolerance to guard against this singularity problem. A parameter estimate is not changed at an iteration if more than  $1 - \text{TOL}$  proportion of the sum of squares of partial derivatives with respect to that parameter can be expressed with partial derivatives of other parameters.

**Loss convergence.** When the relative improvement in the loss function for an iteration is less than the specified value, SYSTAT declares that a solution has been found. Note that, for convergence, both loss convergence and parameter convergence must be satisfied.

**Parameter.** When the largest relative improvement of parameters for an iteration is less than the specified value, SYSTAT considers that the estimates of the parameters have converged. Each parameter estimate must satisfy this criterion.

**Fix.** Specify names of parameters to be held fixed at a constant value. SYSTAT estimates the remaining parameters and tests whether the result differs from that for the full model. An example is  $p3 = 1.0$ .

**Use Marquardt.** The Marquardt method of inflating the diagonal of the (Jacobian/Jacobian) matrix by  $n$ . This speeds convergence when initial values are far from the estimates and when the estimates of the parameters are highly intercorrelated. This method is similar to "ridging," except that the inflation factor  $n$  is omitted from final iterations.

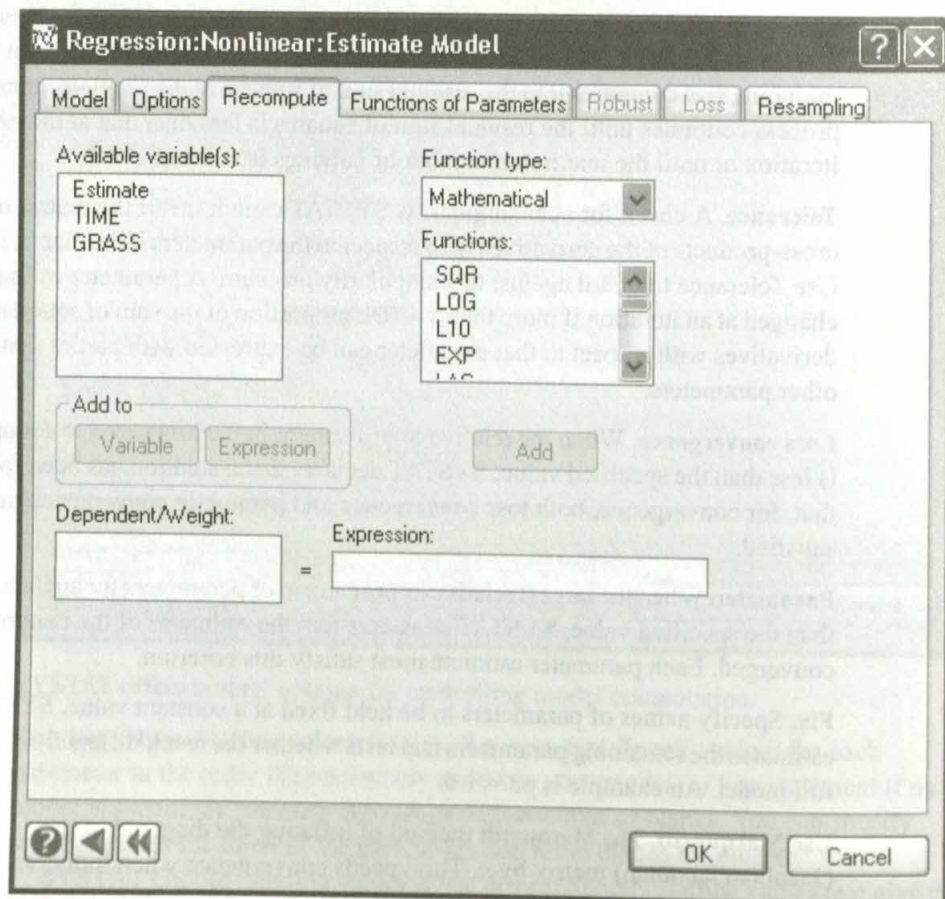
**Mean square error scale.** Rescales the mean square error to 1 at the end of the iterations.



### Recompute

The dependent variable or the weight variable can be recomputed after each iteration, using the current values of the parameters.

You can invoke the recompute option by clicking the Recompute tab in the Nonlinear Regression: Estimate Model dialog box.

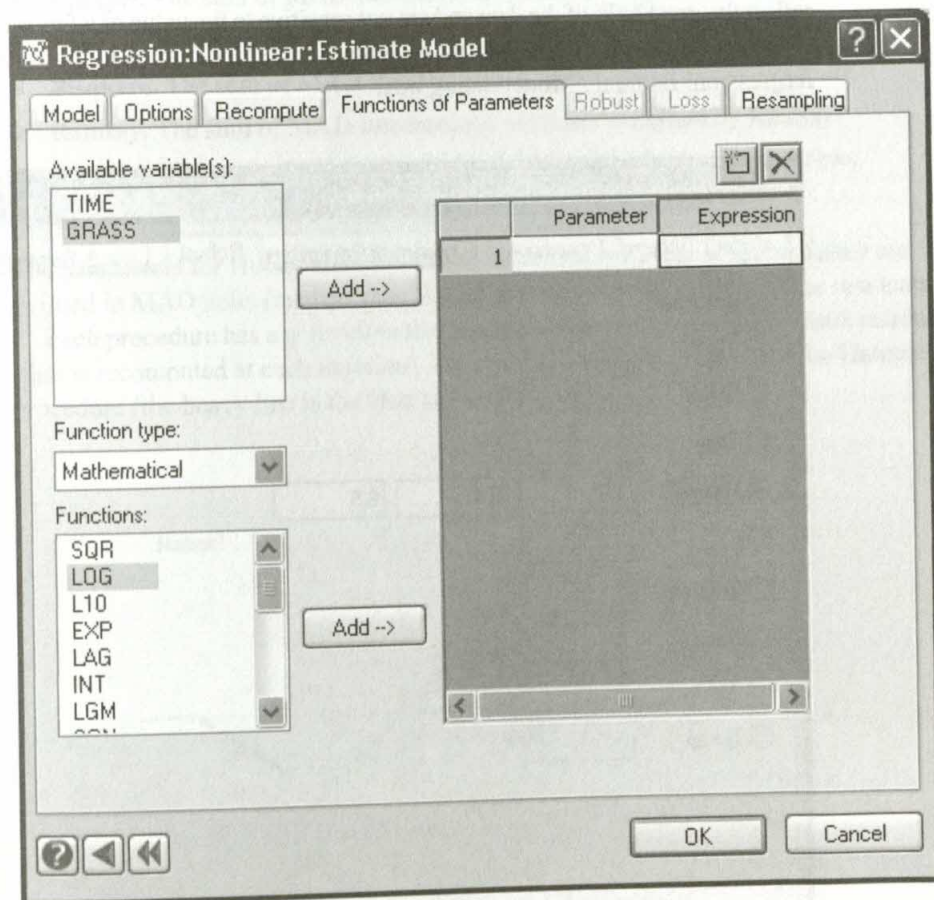


Select an appropriate variable as Dependent/Weight variable from the list of Available variable(s) by clicking the Add button. If you want to use a function in your expression, choose a Function type from the drop-down list, select the function in the functions list, and click Add.



### Functions of Parameters

To specify the function of the parameter click Functions of Parameters tab.



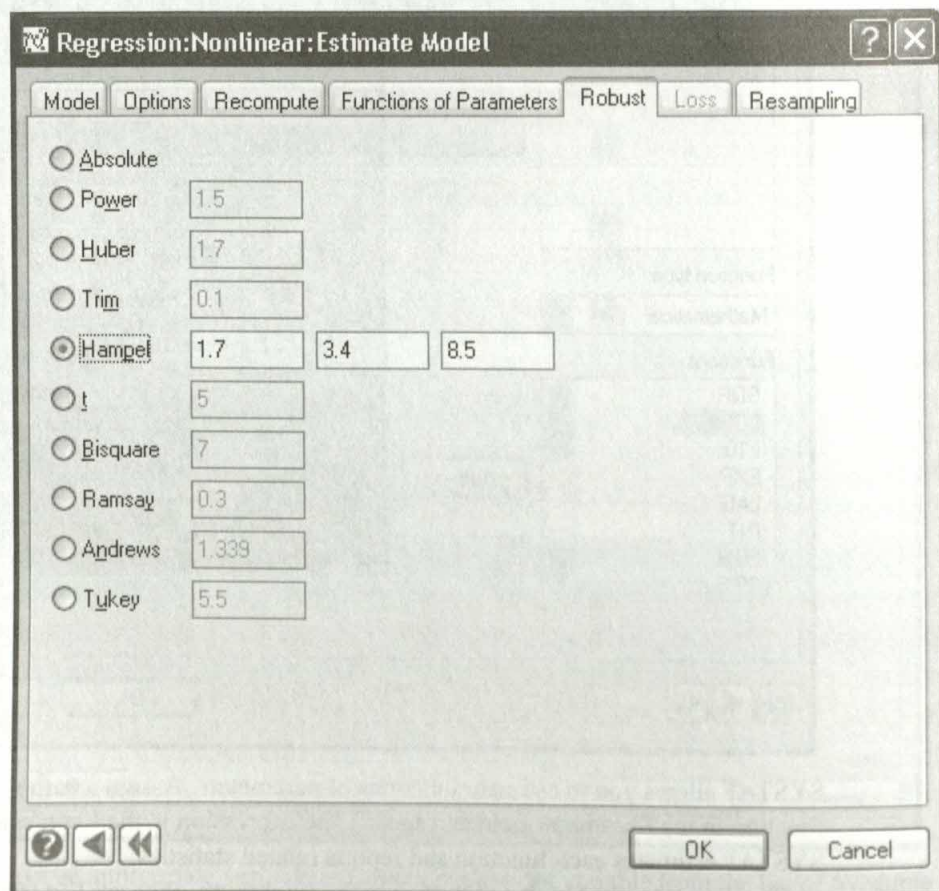
SYSTAT allows you to estimate functions of parameters. Assign a name to each function in the Parameter field and specify the expression in the Expression field. SYSTAT estimates each function and reports related statistics.

If you want to use a built-in function in the expression, choose a Function type from the drop-down list, select the function in the functions list, and click Add.

## Robust

When your dependent variable contains outliers, a robust regression procedure can downweight their influence on the parameter estimates. Thus, the resulting estimates reflect the great bulk of the data and are not sensitive to the value of a few unusual cases.

To specify a robust analysis, select Robust under Estimation in the Nonlinear Regression: Estimate Model dialog box.



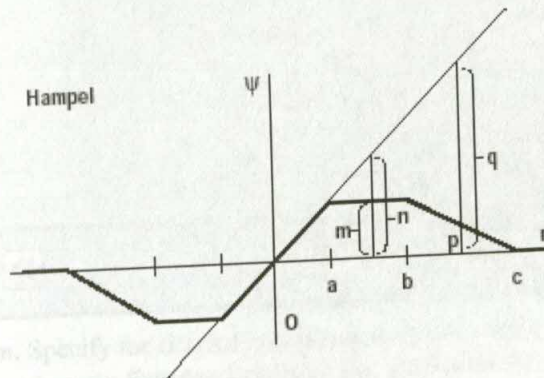
The available methods include:

- **Absolute.** The sum of absolute values of residuals.
- **Power.** The sum of the  $n$ th power of absolute values of residuals.

- **Huber.** The sum of MAD standardized residuals weighted by Huber.
- **Trim.** Trims the  $n$  proportions of the residuals (those with the largest absolute values) and minimizes the sum of squares of the remaining residuals.
- **Hampel.** The sum of MAD standardized residuals weighted by Hampel.
- **$t$ .** A  $t$  distribution with  $df$  (degrees of freedom).
- **Bisquare.** The sum of MAD standardized residuals weighted by Bisquare.
- **Ramsay.** The sum of MAD standardized residuals weighted by Ramsay.
- **Andrews.** The sum of MAD standardized residuals weighted by Andrews.
- **Tukey.** The sum of MAD standardized residuals weighted by Tukey.

The parameters for Huber, Hampel,  $t$ , Bisquare, Ramsay, Andrews, and Tukey are defined in MAD units (median absolute deviations from the median of the residuals).

Each procedure has a  $\psi$  function that is used to construct a weight for each residual (that is recomputed at each iteration). Here is the weighting scheme for the Hampel procedure (the heavy line is the Hampel  $\psi$  function):



for  $| \text{residual} | < a$   
 $a < | \text{residual} | < b$   
 $b < | \text{residual} | < c$   
 $c < | \text{residual} |$

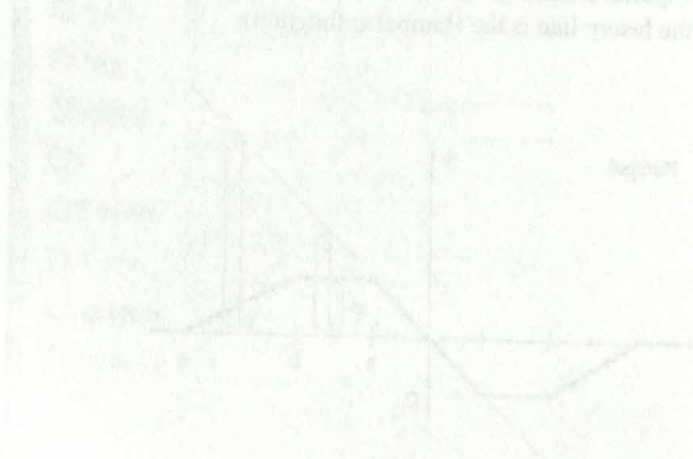
the weight  $((\text{residual})/\text{residual})$  is 1.0  
the weight is  $m/n$   
the weight is  $p/q$   
the weight is 0.0

Nonlinear Model's default values for a, b, and c are 1.7, 3.4, and 8.5, respectively. So, if the size of the residual is less than 1.7, the weight is one; if it is over 8.5, the weight is zero. As the residual increases in absolute value, the weight decreases.

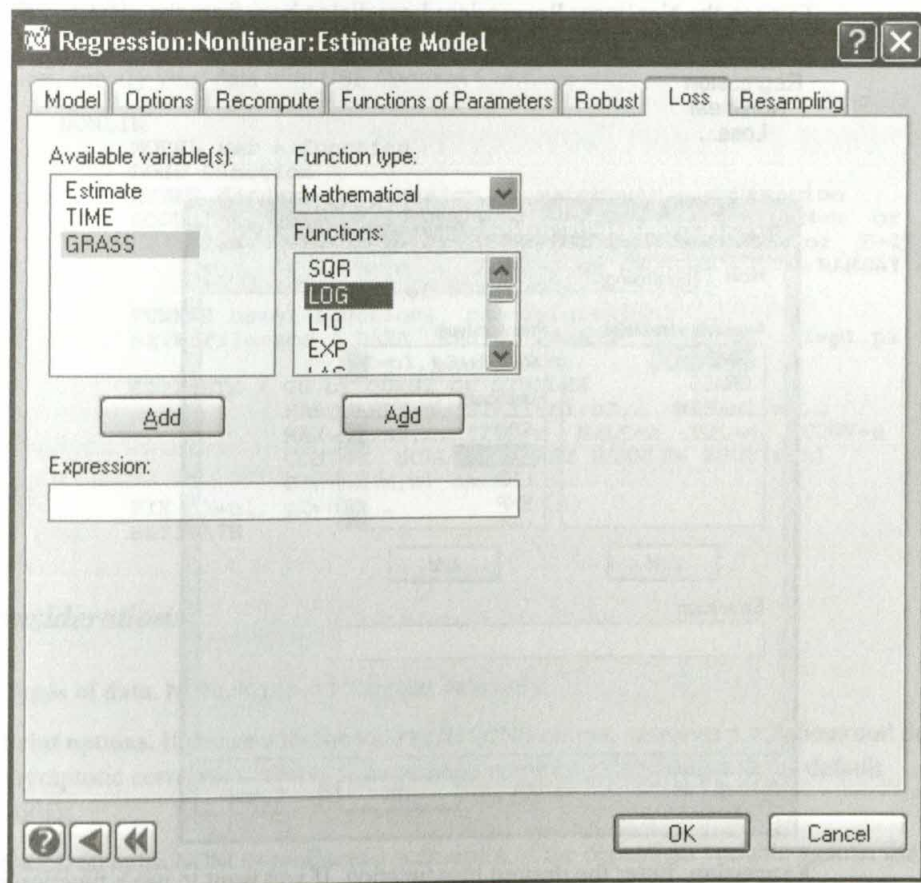
### ***Loss Function for Nonlinear Model Estimation***

As an alternative to least-squares and robust regression, you can specify a custom loss function to apply in model estimation. The default (least-squares) loss function is  $(\text{depvar} - \text{estimate})^2$ . The word "estimate" in the function is the fitted value from your model. It is a special Nonlinear Model word, so you should not name a variable *ESTIMATE*. The model defines the parameters (so new parameters cannot be introduced in the loss function).

To specify a loss function for a model, select Loss function under the Estimation in the Nonlinear Regression: Estimate Model dialog box.







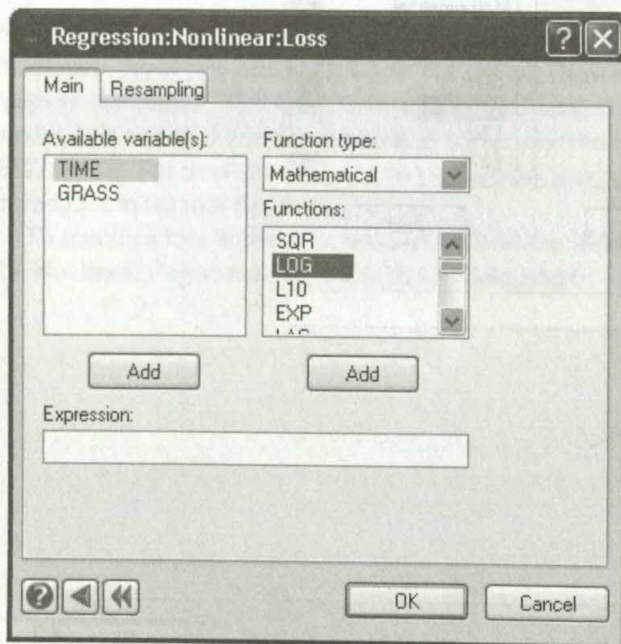
**Expression.** Specify the desired loss function. If you want to use a function in the expression, choose a Function type from the drop-down list, select the function in the functions list, and click Add.

### ***Loss Functions for Analytic Function Minimization***

You can also use nonlinear estimation to minimize an algebraic function. Such a function requires no model specification. As a result, the loss function defines the parameters and SYSTAT computes no *estimates* for a dependent variable.

To open the Nonlinear Regression: Loss dialog box, from the menus choose:

Analyze  
Regression  
Nonlinear  
Loss...



**Expression.** Enter the desired loss function. If you want to use a function in the expression, choose a Function type from the drop-down list, select the function in the functions list, and click Add.

If estimation problems arise, use an alternative estimation method. The Simplex method generally does better with algebraic expressions that incur roundoff error.

## Using Commands

First, specify your data with *USE filename*. Continue with:

```

NONLIN
MODEL var = function
LOSS function
RESET depvar = expression or weightvar = expression
ROBUST argument / ABSOLUTE or POWER=n or TRIM=n or
                    HUBER=n, or HAMPEL=n1,n2,n3 or T=df
                    or BISQUARE=n or ANDREWS = n or RAMSAY =
                    n or TUKEY = n
FUNPAR name1=function1, name2=function2, ...
SAVE filename / DATA RESID PARAMS RS=p1,p2 CI=p1,p2
                    CR=p1,p2 CONFI=n
ESTIMATE / GN or QUASI or SIMPLEX
                    MARQUARDT=n START=n1,n2,... MIN=n1,n2,...
                    MAX=n1,n2,..., ITER=n HALF=n TOL=n LCONV=n
                    CONV=n SCALE RESTART SAMPLE= BOOT(m,n)
                    SIMPLE(m,n) JACK
FIX p1=n1, p2=n2, ...
ESTIMATE

```

## Usage Considerations

**Types of data.** NONLIN uses rectangular data only.

**Print options.** If you specify the PLENGTH LONG output, casewise predictions and the asymptotic correlation matrix of parameters are printed in addition to the default output.

**Quick Graphs.** NONLIN produces a scatterplot of the dependent variable against the variables in the model expression. The fitted function appears as either a line or a surface. If the model expression contains three or more variables, only the first two appear in the plot.

**Saving files.** In nonlinear modeling, you can save residuals, estimated values, and variables from your model statement, parameter values, loss function values surrounding the converged minimum, or data for plotting the Cook-Weisberg confidence intervals or two-parameter confidence region.

**BY groups.** NONLIN produces separate results for each level of any BY variable.

**Case frequencies.** NONLIN uses a FREQUENCY variable, if present, to duplicate cases.

**Case weights.** You can weight cases in NONLIN by specifying a WEIGHT variable.

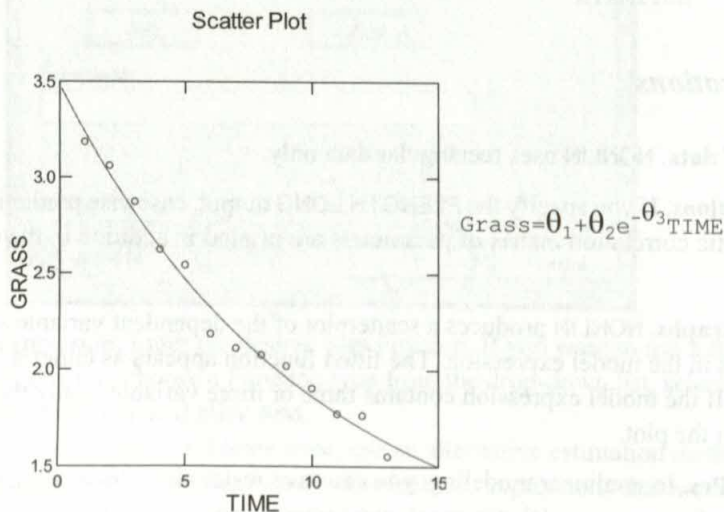
## Examples

### Example 1

#### Nonlinear Model with Three Parameters

For this first example, we do not specify any options specific to NONLIN; we simply specify the model using the operators and functions available for SYSTAT's transformations. Here, we use the default Gauss-Newton algorithm that computes exact derivatives.

The Pattison data are from a 1987 JASA article by G. P. Y. Clarke (Clarke took the data from an unpublished thesis by N. B. Pattinson). For 13 grass samples collected in a pasture, Pattison recorded the number of weeks since grazing began in the pasture (*TIME*) and the weight of grass (*GRASS*) cut from 10 randomly sited quadrants. He then fit the Mitcherlitz equation. Here is the model with the Quick Graph from its fit:



The input is:

```
USE PATTISON
NONLIN
  PLENGTH LONG
  MODEL GRASS = p1 + p2*EXP(-p3*TIME)
  ESTIMATE
```



The output is:

#### Iteration History

No.	Loss	P1	P2	P3
0	22.082	1.010	1.020	1.030
1	12.061	1.170	0.183	-0.153
2	11.247	1.722	-0.053	-0.212
3	5.301	2.727	-0.315	0.112
4	2.817	0.971	2.510	0.186
5	0.128	1.209	2.235	0.109
6	0.054	0.967	2.515	0.102
7	0.053	0.963	2.519	0.103
8	0.053	0.963	2.519	0.103
9	0.053	0.963	2.519	0.103

Dependent Variable : GRASS

#### Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	70.871	3	23.624
Residual	0.053	10	0.005
Total	70.925	13	
Mean corrected	3.309	12	

#### R-squares

Raw R-square (1-Residual/Total) : 0.999  
 Mean Corrected R-square (1-Residual/Corrected) : 0.984  
 R-square (Observed vs Predicted) : 0.984

#### Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
P1	0.963	0.322	2.995	0.247	1.680
P2	2.519	0.266	9.478	1.927	3.111
P3	0.103	0.026	4.041	0.046	0.160

#### Residuals

Case	GRASS Observed	GRASS Predicted	Residual
1	3.183	3.235	-0.052
2	3.059	3.013	0.046
3	2.871	2.812	0.059
4	2.622	2.631	-0.009
5	2.541	2.468	0.073
6	2.184	2.320	-0.136
7	2.110	2.188	-0.078
8	2.075	2.068	0.007
9	2.018	1.959	0.059
10	1.903	1.862	0.041
11	1.770	1.774	-0.004
12	1.762	1.695	0.067
13	1.550	1.623	-0.073

#### Asymptotic Correlation Matrix of Parameters

	p1	p2	p3
p1	1.000		
p2	-0.972	1.000	
p3	0.984	-0.923	1.000

The estimates of parameters converged in nine iterations. At each iteration, Nonlinear Model prints the number of the iteration, the loss, or the residual sum of squares (RSS), and the estimates of the parameters. At step 0, the estimates of the parameters are the starting values chosen by SYSTAT or specified by the user with the START option of ESTIMATE. The residual sum of squares is

$$\sum w(y-f)^2$$

where  $y$  is the observed value,  $f$  is the estimated value, and  $w$  is the value of the case weight (its default is 1.0).

Sums of squares (SS) appearing in the output include:

Regression:  $\sum wy^2 - \sum w(y-f)^2$

Residual:  $\sum w(y-f)^2$

Total:  $\sum wy^2$

Mean corrected:  $\sum w(y-\bar{y})^2$

The Raw  $R^2$  (*Regression SS / Total SS*) is the proportion of the variation in  $y$  that is explained by the sum of squares due to regression. Some researchers object to this measure because the means are not removed. The Mean corrected  $R^2$  tries to adjust for this. Many researchers prefer the last measure of  $R^2$  (*observed vs. predicted squared*). It is the correlation squared between the observed values and the predicted values.

A period (there is none here) for the asymptotic standard error indicates a problem with the estimate (the correlations among the estimated parameters may be very high, or the value of the function may not be affected if the estimate is changed). Read Parameter/ASE, the estimate of each parameter divided by its asymptotic standard error, roughly as a  $t$  statistic.

The Wald Confidence Intervals for the estimates are defined as  $EST \pm t * ASE$  for the  $t$  distribution with residual degrees of freedom ( $df = 10$  in this example). SYSTAT prints the 95% confidence intervals. Use `CONF=n` to specify a different confidence level.

SYSTAT computes asymptotic standard errors and correlations by estimating the  $INV(J'J)$  matrix after iterations have terminated. The matrix is computed from the asymptotic covariance matrix that inverts  $INV(J'J) * RMS$ , where  $J$  is the Jacobian and  $RMS$  is the residual mean squared. You should examine your model for redundant

parameters. If the JJ matrix is singular (parameters are very highly intercorrelated), SYSTAT prints a period to mark parameters with problems. In this example, the parameters are highly intercorrelated; the model may be overparameterized.

## Example 2

### Confidence Curves and Regions

Confidence curves and regions provide information about the certainty of your parameter estimates. The usual Wald confidence intervals can be misleading when intercorrelations among the parameters are high.

**Confidence curves.** Cook and Weisberg construct confidence curves by plotting an assortment of potential estimates of a specific parameter on the  $y$  axis against the absolute value of a  $t$  statistic derived from the residual sum of squares (RSS) associated with each parameter estimate. To obtain the values for the  $x$  axis, SYSTAT:

- Computes the model as usual and saves RSS.
- Fixes the value of the parameter of interest (of, for example, the estimate plus half the standard error of the estimate), recomputes the model, and saves RSS\*.
- Computes the  $t$  statistic:

$$t = \frac{\sqrt{\frac{\text{RSS}^* - \text{RSS}}{1}}}{\sqrt{\frac{\text{RSS}}{n - p}}}$$

- Repeats the above steps for other estimates of the parameter.

Now SYSTAT plots each parameter estimate against the absolute value of its associated  $t^*$  statistic. Vertical lines at the 90, 95, and 99 percentage points of the  $t$  distribution with  $(n - p)$  degrees of freedom provide a useful frequentist calibration of the plot.

To illustrate the usefulness of confidence curves, we again use the Pattison data used in the three-parameter nonlinear model example. Recall that the parameter estimates were:

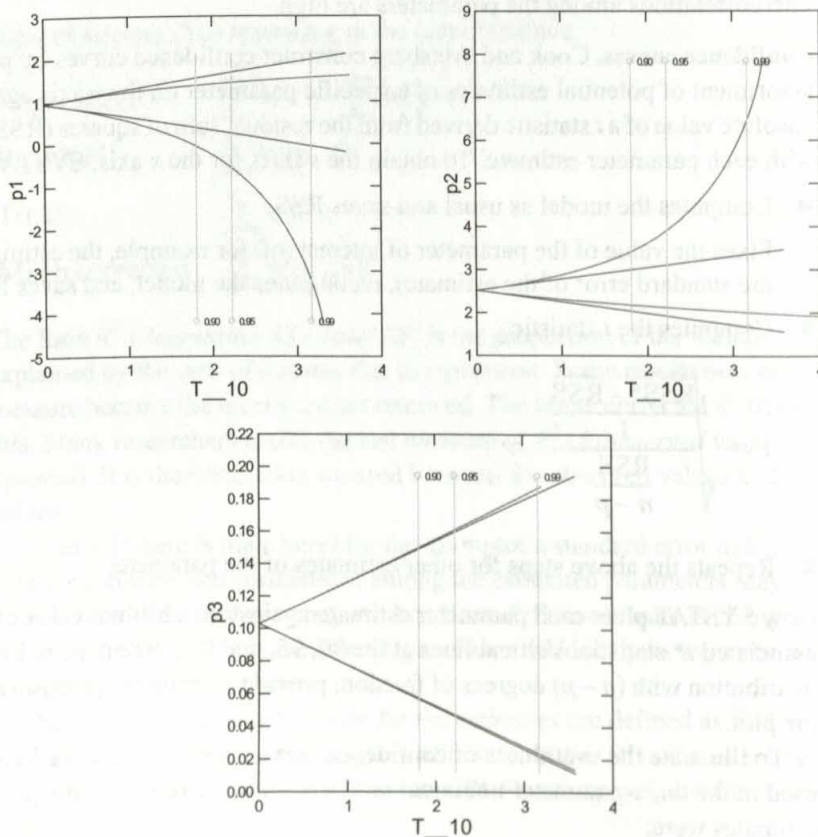
$$\begin{aligned} p1 &= 0.93 \\ p2 &= 2.519 \\ p3 &= 0.103 \end{aligned}$$

To produce the Cook-Weisberg confidence curves for the model,

the input is:

```
USE PATTISON
NONLIN
  MODEL GRASS = p1 + p2*EXP(-p3*TIME)
  SAVE PATTICI / CI=p1, p2, p3
  ESTIMATE
  SUBMIT '&SAVE\PATTICI'
```

The output is:



The nonvertical straight lines (blue on a computer monitor) are the Wald 95% confidence intervals and the solid curves are the Cook-Weisberg confidence curves. The vertical lines show the 90th, 95th, and 99th percentiles of the  $t$  distribution with  $n - p = 10$  degrees of freedom.



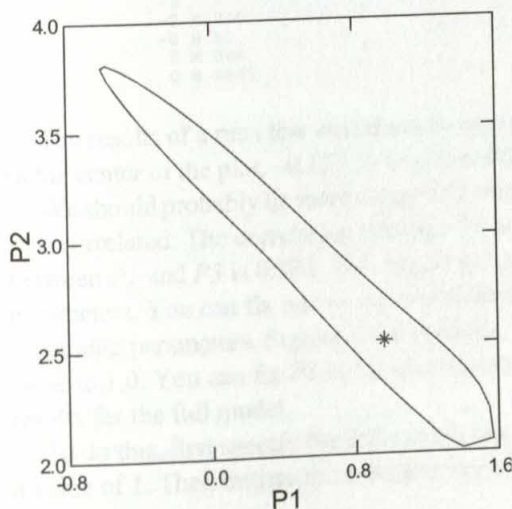
For  $P1$  and  $P2$ , the coverage of the Wald intervals differs markedly from that of the Cook-Weisberg (C-W) curves. The 95% interval for  $P1$  on the C-W curve is approximately from  $-0.58$  to  $1.45$ ; the Wald interval extends from  $0.247$  to  $1.68$ . The steeply descending lower C-W curve indicates greater uncertainty for smaller estimates of  $P1$ . For  $P2$ , the C-W interval ranges from  $2.12$  to  $3.92$ ; the Wald interval ranges from  $1.9$  to  $3.1$ . The agreement between the two methods is better for  $P3$ . The C-W curves show that the distributions of estimates for  $P1$  and  $P2$  are quite asymmetric.

**Confidence region.** SYSTAT also provides the CR option for confidence regions. When there are more than two parameters in the model, this feature causes Nonlinear Model to search for the best values of the additional parameters for each combination of estimates for the first two parameters.

The input is:

```
USE PATTISON
NONLIN
MODEL GRASS = p1 + p2*EXP(-p3*TIME)
SAVE PATTTCR / CR=p1, p2
ESTIMATE
SUBMIT '&SAVE\ PATTTCR'
```

The output is:



You can also specify the level of confidence. For example,

```
SAVE PATTTCR / CR=p1, p2 CONFI=.90
```

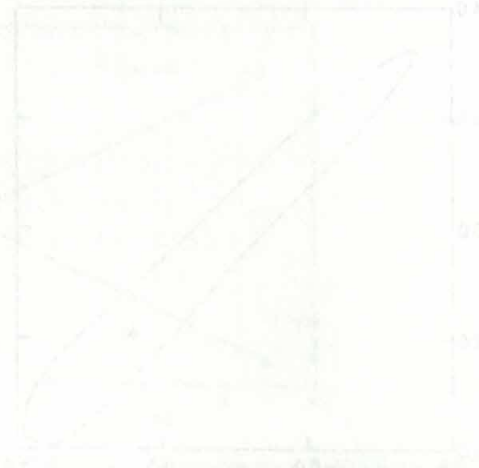
### Example 3

#### Fixing Parameters and Evaluating Fit

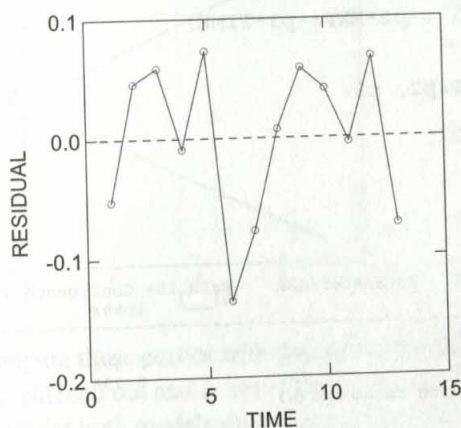
In the three-parameter nonlinear model example, the  $R^2$  between the observed and predicted values is 0.984, indicating good agreement between the data and fitted values. However, there may be consecutive points across time where the fitted values are consistently overestimated or underestimated. We can look for trends in the residuals by plotting them versus *TIME* and connecting the points with a line. A stem-and-leaf plot will tell us if extreme values are identified as outliers (outside values or far outside values).

The input is:

```
USE PATTISON
NONLIN
MODEL GRASS = p1 + p2*EXP(-p3*TIME)
SAVE MYRESIDS / DATA
ESTIMATE
USE MYRESIDS
PLOT RESIDUAL*TIME / LINE YLIMIT=0
STEM RESIDUAL
```



The output is:



Stem and Leaf Plot of Variable: RESIDUAL, N = 13

```

Minimum      : -0.136
Lower Hinge  : -0.052
Median       : 0.007
Upper Hinge  : 0.059
Maximum      : 0.073

-1      3
-0 H 775
-0 H 00
0 M 044
0 H 5567

```

The results of a runs test would not be significant here. The large negative residual in the center of the plot,  $-0.137$ , is not identified as an outlier in the stem-and-leaf plot.

We should probably be more concerned about the fact that the parameters are highly intercorrelated: The correlation between  $P1$  and  $P2$  is  $-0.972$ , and the correlation between  $P1$  and  $P3$  is  $0.984$ . This might indicate that our model has too many parameters. You can fix one or more parameters and let SYSTAT estimate the remaining parameters. Suppose, for example, that similar studies report a value of  $P1$  close to  $1.0$ . You can fix  $P1$  at  $1.0$  and then test whether the results differ from the results for the full model.

To do this, first specify the full model. Use FIX to specify the parameter as  $P1$  with a value of  $1$ . Then initiate the estimation process.

The input is:

```
USE PATTISON
NONLIN
MODEL GRASS = p1 + p2*EXP(-p3*TIME)
ESTIMATE
FIX p1=1
SAVE PATTICI / CI=p2, p3
ESTIMATE
SUBMIT '&SAVE\PATTICI'
```

The output is:

#### Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
P1	1.000	0.000	.	.	.
P2	2.490	0.060	41.662	2.358	2.621
P3	0.106	0.004	23.728	0.096	0.116

#### Analysis of the Effect of Fixing Parameter(s)

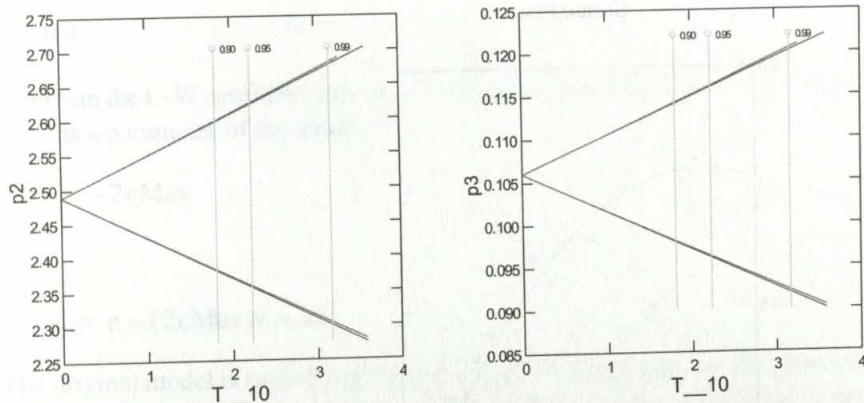
Source	SS	df	Mean Squares	F-ratio	p-value
Fixed Parameter(s)	0.000	1	0.000	0.014	0.908
Residual	0.053	10	0.005		

In the analysis of the effect of fixing parameter(s), F test tests the hypothesis that  $P1=1$ . In our output,  $F = 0.014$  ( $p\text{-value} = 0.908$ ), indicating that there is no significant difference between the two models. This is not surprising, considering the similarity of the results:

	Three parameters	P1 fixed at 1.0
P1	0.963	1.000
P2	2.519	2.490
P3	0.103	0.106
RSS	0.053	0.054
$R^2$	0.984	0.984

There are some differences between the two models. The correlation between  $P2$  and  $P3$  is  $-0.923$  for the full model and  $0.810$  when  $P1$  is fixed. The most striking difference is in the Wald intervals for  $P2$  and  $P3$ . When  $P1$  is fixed, the Wald interval for  $P2$  is less than one-fourth of the interval for the full model. The interval for  $P3$  is less than one-fifth the interval for the full model. Let's see what information the C-W curves provide about the uncertainty of the estimates. Here are the curves for the model with  $P1$  fixed:





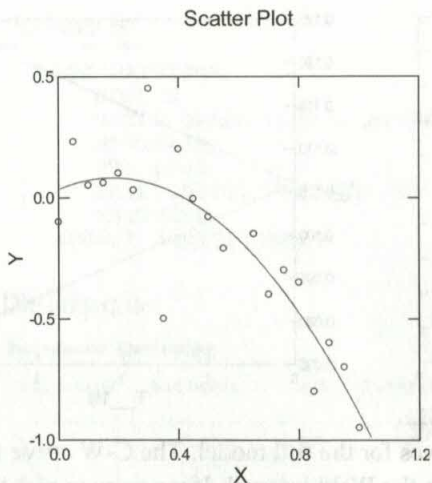
Compare these curves with the curves for the full model. The C-W curve for  $P2$  has straightened out and is very close to the Wald interval. If we were to plot the  $P2$  C-W curve for both models on the same axes, the wedge for the fixed  $P1$  model would be only a small slice of the wedge for the full model.

#### Example 4

#### Functions of Parameters

Frequently, researchers are not interested in the estimates of the parameters themselves, but instead want to make statements about functions of parameters. For example, in a logistic model, they may want to estimate  $LD50$  and  $LD90$  and determine the variability of these estimates. You can specify functions of parameters in Nonlinear Model. SYSTAT evaluates the function at each iteration and prints the standard error and the Wald interval for the estimate after the last iteration.

We look at a quadratic function described by Cook and Weisberg. Here is the Quick Graph that results from fitting the model:



This function reaches its maximum at  $-b/2c$ . However, for the data given by Cook and Weisberg, this maximum is close to the smallest  $x$ . That is, to the left of the maximum, there is little of the response curve.

In SYSTAT, you can estimate the maximum (and get Wald intervals) directly from the original quadratic by using FUNPAR.

The input is:

```
USE QUAD
NONLIN
MODEL y = a + b*x + c*x^2
FUNPAR MAX = -b/(2*c)
ESTIMATE
```

The output is:

#### Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
A	0.034	0.117	0.292	-0.213	0.282
B	0.524	0.555	0.944	-0.647	1.694
C	-1.452	0.534	-2.718	-2.579	-0.325
MAX	0.180	0.128	1.409	-0.090	0.450

Using the Wald interval, we estimate that the maximum response occurs for an  $x$  value between  $-0.09$  and  $0.45$ .

### C-W Curves

To obtain the C-W confidence curves for *MAX*, we have to re-express the model so that *MAX* is a parameter of the model:

$$b = -2c\text{MAX}$$

so

$$y = a - (2c\text{MAX})x + cx^2$$

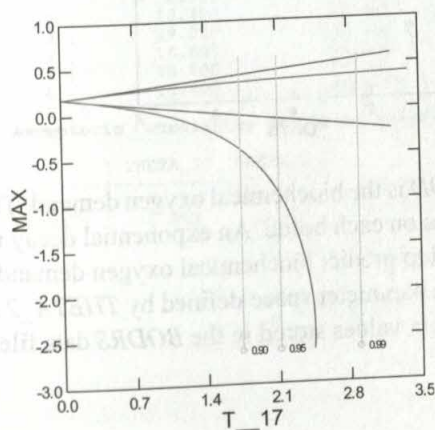
The original model is easy to compute because it is linear. The reparameterized model is not as well-behaved, so we use estimates from the first run as starting values and request C-W confidence curves.

The input is:

```
MODEL y=a - (2*c*MAX)*x + c*x^2
SAVE QUADCW / CI=MAX
ESTIMATE / START=0.034, -1.452, 0.180
SUBMIT '&SAVE\QUADCW'
```

The C-W confidence curves describe our uncertainty about the *x* value at which the expected response is maximized much better than the Wald interval does.

The output is:



The picture provides clear information about the *MAX* response in the positive direction. We can be confident that the value is less than 0.4 because the C-W curve is lower than the Wald interval on the 95th percentile line. The lower bound is much less clear; it could certainly be lower than the Wald interval indicates.

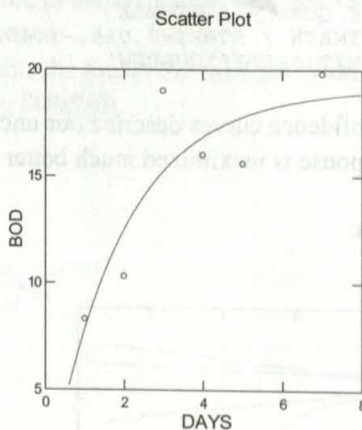
### Example 5

#### Contouring the Loss Function

You can save loss function values along contour curves and then plot the loss function. For this example, we use the *BOD* data (Bates and Watts, 1988). These data were taken from stream samples in 1967 by Marske. Each sample bottle was inoculated with a mixed culture of microorganisms, sealed, incubated, and opened periodically for analysis of dissolved oxygen concentration.

The data are:

DAYS	BOD
1.0	8.3
2.0	10.3
3.0	19.0
4.0	16.0
5.0	15.0
7.0	19.8

$$BOD = \theta_1(1 - e^{-\theta_2 \text{DAYS}})$$


where *DAYS* is time in days and *BOD* is the biochemical oxygen demand. The six *BOD* values are averages of two analyses on each bottle. An exponential decay model with a fixed rate constant was estimated to predict biochemical oxygen demand.

Let's look at the contours of the parameter space defined by *THETA\_2* with *THETA\_1*. We use loss function data values stored in the *BODRS* data file.



The input is:

```
USE BOD
NONLIN
MODEL BOD = theta_1*(1-EXP(-theta_2*DAYS))
PLENGTH LONG
SAVE BODRS / RS
ESTIMATE
SUBMIT '&SAVE\BODRS'
```

The output is:

Dependent Variable :BOD

#### Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	1401.390	2	700.695
Residual	25.990	4	6.498
Total	1427.380	6	
Mean corrected	107.213	5	

#### R-squares

Raw R-square (1-Residual/Total) : 0.982  
 Mean Corrected R-square (1-Residual/Corrected) : 0.758  
 R-square(Observed vs Predicted) : 0.758

#### Parameter Estimates

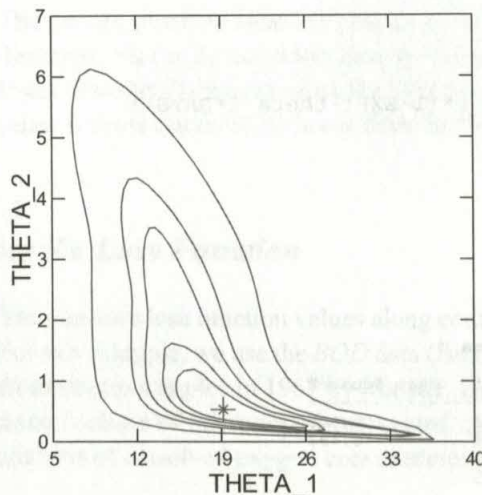
Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
THETA_1	19.143	2.496	7.670	12.213	26.072
THETA_2	0.531	0.203	2.615	-0.033	1.095

#### Residuals

Case	BOD Observed	BOD Predicted	Residual
1	8.300	7.887	0.413
2	10.300	12.525	-2.225
3	19.000	15.252	3.748
4	16.000	16.855	-0.855
5	15.600	17.797	-2.197
6	19.800	18.678	1.122

#### Asymptotic Correlation Matrix of Parameters

	THETA_1	THETA_2
THETA_1	1.000	
THETA_2	-0.853	1.000



The kidney-shaped area near the center of the plot is the region where the loss function is minimized. Any parameter value combination (that is, any point inside the kidney) produces approximately the same loss function.

### Example 6 Maximum Likelihood Estimation

Because NONLIN includes a loss function, you can maximize the likelihood of a function in the model equation. The way to do this is to minimize the negative of the log-likelihood.

Here is an example using the *IRIS* data. Let's compute the maximum likelihood estimates of the mean and variance of *SEPALWID* assuming a normal distribution for the first species in the *IRIS* data. For a sample of  $n$  independent normal random variables, the log-likelihood function is:

$$L(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi) - \frac{n}{2} \ln(\sigma^2) - \frac{1}{2\sigma^2} \sum (X - \mu)^2$$

However, we can use the ZDF function as a shortcut. In this example, we minimize the negative of the log-likelihood with LOSS and thus maximize the likelihood. SYSTAT's small default starting values for *MEAN* and *SIGMA* (0.101 and 0.100) will produce very large  $z$  scores  $((x - \text{mean}) / \text{sigma})$  and values of the density close to 0, so we

arbitrarily select larger starting values. We use the *IRIS* data. Under **SELECT**, we specify **SPECIES = 1**. Then, we type in our **LOSS** statement. Finally, we use **ESTIMATE**'s **START** option to specify start values (2,2).

The input is:

```
USE IRIS
NONLIN
SELECT SPECIES=1
LOSS -log(ZDF(SEPALWID,MEAN,SIGMA))
ESTIMATE / START=2,2
```

The output is:

**Parameter Estimates**

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
MEAN	3.428	0.053	65.255	3.322	3.534
SIGMA	0.375	0.037	10.102	0.301	0.450

Note that the least-squares estimate of sigma (0.379) computed using **CSTATISTICS** is larger than the biased maximum likelihood estimate here (0.375).

### Example 7

#### *Iteratively Reweighted Least-Squares for Logistic Models*

Cox and Snell (1989) report the following data on tests among objects for failures after certain times. These data are in the *COX* data file—*FAILURE* is the number of failures and *COUNT* is the total number of tests.

Cox uses a logistic model to fit the failures:

$$\text{estimate} = (\text{count}) \frac{e^{\beta_0 + \beta_1 \text{time}}}{1 + e^{\beta_0 + \beta_1 \text{time}}}$$

The log-likelihood function for the logit model is:

$$L(\beta_0, \beta_1) = \sum [p \ln(\text{estimate}) + (1 - p) \ln(1 - \text{estimate})]$$

where the sum is over all observations. Because the counts differ at each time, the variances of the failures also differ. If *FAILURE* is randomly sampled from a binomial, then,

$$\text{VAR}(\text{failure}) = \text{estimate} * (\text{count} - \text{estimate}) / \text{count}$$

Therefore, the weight is  $1/\text{variance}$  :

$$w_i = \text{count} / (\text{estimate} * (\text{count} - \text{estimate}))$$

We use these variances to weight each case in the estimation. On each iteration, the variances are recalculated from the new estimates and used anew in computing the weighted loss function.

In the following commands, we use RESET to recompute the weight after each iteration. The SCALE option of ESTIMATE rescales the mean square error to 1 at the end of the iterations.

The input is:

```
USE COX
NONLIN
  PLENGTH LONG
  LET w = 1
  WEIGHT w
  MODEL FAILURE = COUNT*EXP(-b0-b1*TIME)/,
                (1 + EXP(-b0-b1*TIME))
  RESET W = COUNT / (ESTIMATE*(COUNT-ESTIMATE))
  ESTIMATE / SCALE
```

The output is:

#### Iteration History

No.	Loss	B0	B1
0	162.222	0.101	0.102
1	16.178	2.723	-0.011
2	3.254	4.196	-0.051
3	0.754	5.106	-0.074
4	0.666	5.391	-0.080
5	0.675	5.415	-0.081
6	0.675	5.415	-0.081

Dependent Variable : FAILURE

#### Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	13.038	2	6.519
Residual	0.675	2	0.337
Total	13.712	4	
Mean corrected	10.539	3	



## R-squares

Raw R-square (1-Residual/Total) : 0.951  
 Mean Corrected R-square (1-Residual/Corrected) : 0.936  
 R-square(Observed vs Predicted) : 0.988

Standard Errors of Parameters are rescaled

## Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald	95% Confidence Interval
					Lower Upper
B0	5.415	0.728	7.443		3.989 6.841
B1	-0.081	0.022	-3.610		-0.125 -0.037

## Residuals

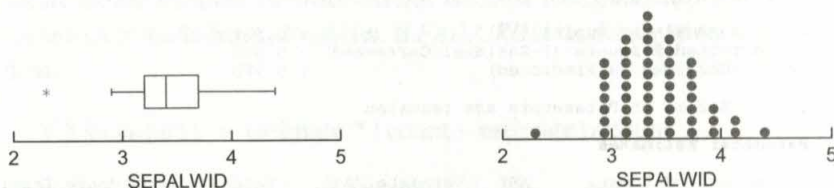
Case	FAILURE Observed	FAILURE Predicted	Residual	Case Weight
1	0.000	0.427	-0.427	2.360
2	2.000	2.132	-0.132	0.475
3	7.000	6.013	0.987	0.173
4	3.000	3.427	-0.427	0.371

Jennrich and Moore (1975) show that this method can be used for maximum likelihood estimation of parameters from a distribution in the exponential family.

**Example 8****Robust Estimation (Measures of Location)**

Robust estimators provide methods other than the mean, median, or mode to estimate the center of a distribution. The sample mean is the least-squares estimate of location; that is, it is the point at which the squared deviations of the sample values are at a minimum. (The sample medians minimize absolute deviations instead of squared deviations.) In terms of  $\psi$  weights, the usual mean assigns a weight of 1.0 to each observation, while the robust methods assign smaller weights to residuals far from the center.

In this example, we use sepal width of the Setosa iris flowers and SELECT SPECIES = 1. We request the usual sample mean and then ask for a 10% trimmed mean, a Hampel estimator, and the median. But first, let's view the distribution graphically. Here is a box-and-whisker display together with a dit plot of the data.



Except for the outlier at the left, the distribution of *SEPALWID* is slightly right-skewed.

### Mean

In the maximum likelihood example, we requested maximum likelihood estimates of the mean and standard deviation.

The input is:

```
USE IRIS
NONLIN
  SELECT SPECIES = 1
  MODEL SEPALWID = MEAN
  ESTIMATE
```

The output is:

#### Iteration History

No.	Loss	MEAN
0	299.377	1.010
1	7.041	3.428
2	7.041	3.428
3	7.041	3.428

Dependent Variable :SEPALWID

#### Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	587.559	1	587.559
Residual	7.041	49	0.144
Total	594.600	50	
Mean corrected	7.041	49	

#### R-squares

```
Raw R-square (1-Residual/Total)      : 0.988
Mean Corrected R-square (1-Residual/Corrected) : 0.000
R-square(Observed vs Predicted)      : 0.000
```

## Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval
				Lower Upper
MEAN	3.428	0.054	63.946	3.320 3.536

**Trimmed Mean**

We enter the following commands after viewing the results for the mean. Note that SYSTAT resets the starting values to their defaults when a new model is specified. If MODEL is not given, SYSTAT uses the final values from the last calculation as starting values for the current task.

For this trimmed mean estimate, SYSTAT deletes the five cases ( $0.1 * 50 = 5$ ) with the most extreme residuals.

The input is:

```
MODEL SEPALWID = TRIMMEAN
ROBUST TRIM = 0.1
ESTIMATE
```

The output is:

## Iteration History

No.	Loss	TRIMMEAN
0	560.487	0.101
1	7.041	3.428
2	3.449	3.428
3	3.372	3.387
4	3.372	3.387
5	3.372	3.387

TRIM Robust Regression

45 cases have positive psi-weights.

The Average Psi-weight : 1.00000

Dependent Variable : SEPALWID

Zero weights, missing data or estimates reduced degrees of freedom

## Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	587.474	1	587.474
Residual	7.126	44	0.162
Total	594.600	45	
Mean corrected	7.041	44	

## R-squares

Raw R-square (1-Residual/Total) : 0.988  
 Mean Corrected R-square (1-Residual/Corrected) : 0.000  
 R-square(Observed vs Predicted) : 0.000

## Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
TRIMMEAN	3.387	0.060	56.451	3.266	3.508

The trimmed estimate deletes the outlier, plus the four flowers on the right side of the distribution with width equal to or greater than 4.0 (if you select the LONG mode of output, you would see that these flowers have the largest residuals).

**Hampel**

We now request a Hampel estimator using the default values for its parameters.

The input is:

```
MODEL SEPALWID = HAMP_EST
ROBUST HAMPEL
ESTIMATE
```

The output is:

## Iteration History

No.	Loss	HAMP_EST
0	560.487	0.101
1	7.041	3.428
2	5.092	3.428
3	5.072	3.416
4	5.069	3.415
5	5.068	3.414
6	5.068	3.414
7	5.068	3.414
8	5.068	3.414

HAMPEL Robust Regression

50 cases have positive psi-weights.  
The Average Psi-weight : 0.94551  
Dependent Variable : SEPALWID

## Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	587.550	1	587.550
Residual	7.050	49	0.144
Total	594.600	50	
Mean corrected	7.041	49	

## R-squares

Raw R-square (1-Residual/Total) : 0.988  
Mean Corrected R-square (1-Residual/Corrected) : 0.000  
R-square(Observed vs Predicted) : 0.000



## Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
HAMP_EST	3.414	0.054	63.648	3.306	3.522

**Median**

We let NONLIN minimize the absolute value of the residuals for an estimate of the median.

The input is:

```
MODEL SEPALWID = MEDIAN
ROBUST ABSOLUTE
ESTIMATE
```

The output is:

## Iteration History

No.	Loss	MEDIAN
0	299.377	1.010
1	14.368	3.428
2	14.299	3.416
3	14.250	3.408
4	14.221	3.404
5	14.208	3.401
6	14.203	3.400
7	14.201	3.400
8	14.200	3.400
9	14.200	3.400
10	14.200	3.400
11	14.200	3.400
12	14.200	3.400
13	14.200	3.400

ABSOLUTE Robust Regression

50 cases have positive psi-weights.  
 The Average Psi-weight :2.41862E+006  
 Dependent Variable :SEPALWID

## Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	587.520	1	587.520
Residual	7.080	49	0.144
Total	594.600	50	
Mean corrected	7.041	49	

R-squares

Raw R-square (1-Residual/Total) : 0.988  
 Mean Corrected R-square (1-Residual/Corrected) : 0.000  
 R-square(Observed vs Predicted) : 0.000

## Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
MEDIAN	3.400				

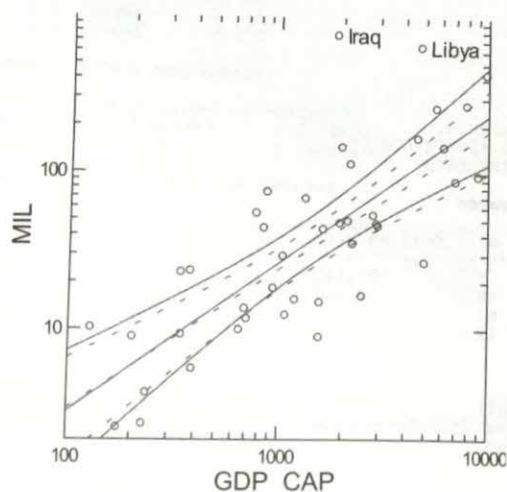
If you request the median for these data in the Basic Statistics procedure, the value is 3.4.

## Example 9 Regression

Usually, you would not use *NONLIN* for linear regression because other procedures are available. If, however, you are concerned about the influence of outliers on the estimates of the coefficients, you should try one of Nonlinear Model's robust procedures.

The example uses the *OURWORLD* data file and we model the relation of military expenditures to gross domestic product using information reported by 57 countries to the United Nations. Each country is a case in our file and *MIL* and *GDP\_CAP* are our two variables. In the transformation example for linear regression, we discovered that both variables require a log transformation, and that Iraq and Libya are outliers.

Here is a scatterplot of the data. The solid line is the least-squares line of best fit for the complete sample (with its corresponding confidence band); the dotted line (and its confidence band) is the regression line after deleting Iraq and Libya from the sample. How do robust lines fit within original confidence bands?



Visually, we see the dotted line-of-best fit falls slightly below the solid line for the complete sample. More striking, however, is the upper curve for the confidence band—the dotted line is considerably lower than the solid one.

We can use NONLIN to fit a least-squares regression line.

The input is:

```
USE OURWORLD
NONLIN
  LET LOG_MIL = L10(MIL)
  LET LOG_GDP = L10(GDP_CAP)
  MODEL LOG_MIL = INTERCEPT + SLOPE*LOG_GDP
  ESTIMATE
```

The output is:

Dependent Variable : LOG\_MIL  
Zero weights, missing data or estimates reduced degrees of freedom

**Sum of Squares and Mean Squares**

Source	SS	df	Mean Squares
Regression	194.332	2	97.166
Residual	6.481	54	0.120
Total	200.813	56	
Mean corrected	24.349	55	

**R-squares**

Raw R-square (1-Residual/Total) : 0.968  
Mean Corrected R-square (1-Residual/Corrected) : 0.734  
R-square(Observed vs Predicted) : 0.734

**Parameter Estimates**

Parameter	Estimate	ASE	Parameter/ASE	Wald	95% Confidence Interval
					Lower Upper
INTERCEPT	-1.308	0.257	-5.091	-1.822	-0.793
SLOPE	0.909	0.075	12.201	0.760	1.058

The estimate of the intercept (-1.308) and the slope (0.909) are the same as those produced by GLM. The residual for Iraq (1.216) is identified as an outlier—its Studentized value is 4.004. Libya's residual is 0.77.

### 1st Power

We now estimate the model using a least absolute values loss function (first power regression). We do not respecify the model, so by default, SYSTAT uses our last estimates as starting values. To avoid this, we specify START without an argument.

The input is:

ROBUST ABSOLUTE  
ESTIMATE / START

The output is:

#### Iteration History

No.	Loss	INTERCEPT	SLOPE
0	119.361	0.101	0.102
1	14.708	-1.308	0.909
2	14.658	-1.352	0.920
3	14.630	-1.381	0.927
4	14.614	-1.402	0.932
5	14.614	-1.404	0.932
6	14.614	-1.406	0.933
7	14.613	-1.409	0.934
8	14.612	-1.412	0.934
9	14.612	-1.416	0.935
10	14.611	-1.420	0.936
11	14.610	-1.425	0.937
12	14.610	-1.429	0.938
13	14.609	-1.434	0.939
14	14.608	-1.438	0.940
15	14.608	-1.442	0.941
16	14.607	-1.445	0.942
17	14.607	-1.446	0.942
18	14.607	-1.447	0.943
19	14.607	-1.447	0.943
20	14.607	-1.447	0.943
21	14.607	-1.447	0.943

#### ABSOLUTE Robust Regression

56 cases have positive psi-weights.

The Average Psi-weight : 4.02107E+013

Dependent Variable : LOG\_MIL

Zero weights, missing data or estimates reduced degrees of freedom

#### Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	194.271	2	97.136
Residual	6.542	54	0.121
Total	200.813	56	
Mean corrected	24.349	55	

#### R-squares

Raw R-square (1-Residual/Total) : 0.967

Mean Corrected R-square (1-Residual/Corrected) : 0.731

R-square(Observed vs Predicted) : 0.734

#### Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval
				Lower Upper
INTERCEPT	-1.447	.	.	.
SLOPE	0.943	.	.	.



**Huber**

For the Hampel estimator, the weights begin to be less than 1.0 after the value of the first parameter (1.7). For this Huber estimate, we let the weight taper off sooner by setting the parameter at 1.5.

The input is:

```
ROBUST HUBER = 1.5
ESTIMATE / START
```

The output is:

**Iteration History**

No.	Loss	INTERCEPT	SLOPE
0	119.361	0.101	0.102
1	6.481	-1.308	0.909
2	4.289	-1.308	0.909
3	4.267	-1.338	0.914
4	4.180	-1.357	0.918
5	4.180	-1.369	0.921
6	4.182	-1.373	0.922
7	4.183	-1.374	0.923
8	4.183	-1.374	0.923
9	4.183	-1.374	0.923
10	4.183	-1.374	0.923
11	4.183	-1.374	0.923
12	4.183	-1.374	0.923
13	4.183	-1.374	0.923

HUBER Robust Regression

56 cases have positive psi-weights.

The Average Psi-weight : 0.92050

Dependent Variable : LOG MIL

Zero weights, missing data or estimates reduced degrees of freedom

**Sum of Squares and Mean Squares**

Source	SS	df	Mean Squares
Regression	194.305	2	97.153
Residual	6.508	54	0.121
Total	200.813	56	
Mean corrected	24.349	55	

R-squares

Raw R-square (1-Residual/Total) : 0.968  
 Mean Corrected R-square (1-Residual/Corrected) : 0.733  
 R-square (Observed vs Predicted) : 0.734

**Parameter Estimates**

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
INTERCEPT	-1.374	0.255	-5.398	-1.885	-0.864
SLOPE	0.923	0.073	12.567	0.775	1.070

**5% Trim**

In the linear regression version of this example, we removed Iraq from the sample by specifying:

```
SELECT mil < 700 or SELECT country$ <> 'Iraq'
```

Here, we ask for 5% trimming ( $0.05 \times 56 = 2.8$  or 2 cases).

The input is:

```
ROBUST TRIM = .05
ESTIMATE / START
```

The output is:

**Iteration History**

No.	Loss	INTERCEPT	SLOPE
0	119.361	0.101	0.102
1	6.481	-1.308	0.909
2	4.406	-1.308	0.909
3	4.333	-1.332	0.905
4	4.333	-1.332	0.905
5	4.333	-1.332	0.905

TRIM Robust Regression

54 cases have positive psi-weights.

The Average Psi-weight : 1.00000

Dependent Variable : LOG\_MIL

Zero weights, missing data or estimates reduced degrees of freedom

**Sum of Squares and Mean Squares**

Source	SS	df	Mean Squares
Regression	194.256	2	97.128
Residual	6.557	52	0.126
Total	200.813	54	
Mean corrected	24.349	53	

R-squares

Raw R-square (1-Residual/Total) : 0.967

Mean Corrected R-square (1-Residual/Corrected) : 0.731

R-square(Observed vs Predicted) : 0.734

**Parameter Estimates**

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
INTERCEPT	-1.332	0.264	-5.049	-1.861	-0.803
SLOPE	0.905	0.077	11.829	0.752	1.059

### Example 10

#### Piecewise Regression

Sometimes we need to fit two different regression functions to the same data. For example, sales of a certain product might be strongly related to quality when advertising budgets are below a certain level—that is, when sales are generated by “word of mouth.” Above this advertising budget level, sales may be less strongly related to quality of goods and more by marketing and advertising factors. In these cases, we can fit different sections of the data with different models. It is easier to combine these into a single model, however.

Here is an example of a quadratic function with a ceiling using data from Gilfoil (1982). This particular study is one of several that show that dialog menu interfaces are preferred by inexperienced computer users and that command based interfaces are preferred by experienced users. The data for one subject are in the file *LEARN*. The variable *SESSION* is the session number and *TASKS* is the number of user-controlled tasks (as opposed to dialog) chosen by the subject during a session.

We fit these data with a quadratic model for earlier sessions and a ceiling for later sessions. We use *NONLIN* to estimate the point where the learning hits this ceiling (at six tasks).

The input is:

```
USE LEARN
NONLIN
  PLENGTH LONG
  MODEL TASKS = b*SESSION^2*(SESSION<KNOWN) +,
                b*KNOWN^2*(SESSION>=KNOWN)
  ESTIMATE
```

Note that the expressions (*SESSION*<*KNOWN* and *SESSION*>=*KNOWN*) control which function is to be used—the quadratic or the horizontal line.

The output is:

#### Iteration History

No.	Loss	B	KNOWN
0	313.871	1.010	1.020
1	207.272	0.505	2.042
2	175.758	0.253	3.119
3	152.604	0.126	4.613
4	122.355	0.045	8.026
5	27.032	0.055	11.272
6	16.137	0.054	10.537
7	14.556	0.062	9.678
8	14.418	0.063	9.659
9	14.418	0.063	9.660
10	14.418	0.063	9.660

Dependent Variable :TASKS

**Sum of Squares and Mean Squares**

Source	SS	df	Mean Squares
Regression	445.582	2	222.791
Residual	14.418	18	0.801
Total	460.000	20	
Mean corrected	140.000	19	

**R-squares**

Raw R-square (1-Residual/Total) : 0.969  
 Mean Corrected R-square (1-Residual/Corrected) : 0.897  
 R-square(Observed vs Predicted) : 0.912

**Parameter Estimates**

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
B	0.063	0.007	8.762	0.048	0.079
KNOWN	9.660	0.594	16.269	8.412	10.907

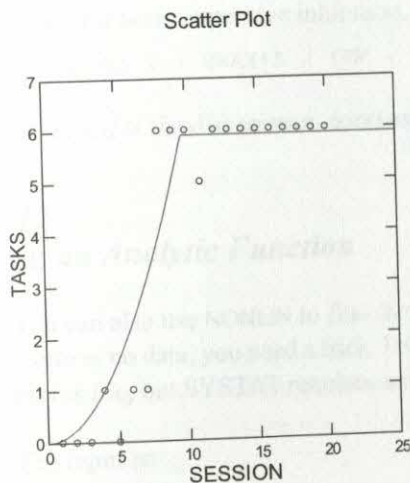
**Residuals**

Case	TASKS Observed	TASKS Predicted	Residual
1	0.000	0.063	-0.063
2	0.000	0.253	-0.253
3	0.000	0.570	-0.570
4	1.000	1.013	-0.013
5	0.000	1.583	-1.583
6	1.000	2.280	-1.280
7	1.000	3.103	-2.103
8	6.000	4.053	1.947
9	6.000	5.130	0.870
10	6.000	5.909	0.091
11	5.000	5.909	-0.909
12	6.000	5.909	0.091
13	6.000	5.909	0.091
14	6.000	5.909	0.091
15	6.000	5.909	0.091
16	6.000	5.909	0.091
17	6.000	5.909	0.091
18	6.000	5.909	0.091
19	6.000	5.909	0.091
20	6.000	5.909	0.091

**Asymptotic Correlation Matrix of Parameters**

	B	KNOWN
B	1.000	
KNOWN	-0.928	1.000





From the Quick Graph, we see that the fit at the lower end is not impressive. We might want to fit a truncated logistic model instead of a quadratic because learning is more often represented with this type of function. This model would have a logistic curve at the lower values of *SESSION* and a flat ceiling line at the upper end. We should use a *LOSS* also to make the maximum likelihood fit.

Piecewise linear regression models with known breakpoints can be fitted similarly. These models look like this:

$$y = b_0 + b_1 * x + b_2 * (x - \text{break}) * (x > \text{break})$$

If the break point is known, then you could also use GLM to do ordinary regression to fit the separate pieces. See Kutner et al. (2004) for an example.

### Example 11

#### Kinetic Models

You can also use *NONLIN* to test kinetic models. The following analysis models competitive inhibition for an enzyme inhibitor. The data are adapted from a conference session on statistical computing with microcomputers (Greco, et al., 1982). We will fit three variables: initial enzyme velocity (*V*), concentration of the substrate (*S*), and concentration of the inhibitor (*I*). The parameters of the model are the maximum velocity (*VMAX*), the Michaelis constant (*KM*) and the dissociation constant of the enzyme-inhibitor complex (*KIS*).

The input is:

```
USE ENZYME
NONLIN
PLENGTH LONG
MODEL V = VMAX*S / (KM*(1 + I/KIS) + S)
ESTIMATE / MIN = 0,0,0
```

The output is:

#### Iteration History

No.	Loss	VMAX	KM	KIS
0	3.568	1.010	1.020	1.030
1	2.289	1.008	0.933	0.000
2	2.286	1.008	0.933	0.000
3	2.082	1.020	0.927	0.001
4	0.027	1.256	0.818	0.023
5	0.014	1.258	0.845	0.027
6	0.014	1.259	0.847	0.027
7	0.014	1.260	0.847	0.027
8	0.014	1.260	0.847	0.027

Dependent Variable :V

#### Sum of Squares and Mean Squares

Source	SS	df	Mean Squares
Regression	15.404	3	5.135
Residual	0.014	43	0.000
Total	15.418	46	
Mean corrected	5.763	45	

#### R-squares

Raw R-square (1-Residual/Total) : 0.999  
 Mean Corrected R-square (1-Residual/Corrected) : 0.998  
 R-square(Observed vs Predicted) : 0.998

#### Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval
				Lower Upper
VMAX	1.260	0.012	104.191	1.235 1.284
KM	0.847	0.027	31.876	0.793 0.900
KIS	0.027	0.001	31.033	0.025 0.029

You could try alternative models for these data such as one for uncompetitive inhibition,

$$\text{MODEL } V = \text{VMAX} * S / (\text{KM} + S + S * I / \text{KII})$$

or one for noncompetitive inhibition,

$$\text{MODEL } V = \text{VMAX} * S / (\text{KM} + \text{KM}/\text{KIS} + S + S * I / \text{KII})$$

where  $KII$  is the dissociation constant of the enzyme-inhibitor-substrate complex.

### Example 12

#### Minimizing an Analytic Function

You can also use NONLIN to find the minimum of an algebraic function. Since this requires no data, you need a trick. Use any data file. We do not use any of the variables in this file, but SYSTAT requires a data file to be open to do a nonlinear estimation.

The input is:

```
USE DOSE
NONLIN
LOSS 100*(U-V^2)^2+(1-V)^2
ESTIMATE / SIMPLEX
```

This particular function is from Rosenbrock (1960). We are using SIMPLEX to save space and because it generally does better with algebraic expressions which incur roundoff error.

The output is:

#### Iteration History

No.	Loss	U	V
0	1.021	1.010	1.020
1	0.931	1.262	1.126
2	0.002	1.005	1.003
3	0.000	0.999	1.000
4	0.000	1.000	1.000
5	0.000	1.000	1.000
6	0.000	1.000	1.000
7	0.000	1.000	1.000
8	0.000	1.000	1.000
9	0.000	1.000	1.000
10	0.000	1.000	1.000

Final Value of Loss Function: 0.000

#### Parameter Estimates

Parameter	Estimate	ASE	Parameter/ASE	Wald 95% Confidence Interval	
				Lower	Upper
U	1.000	.	.	.	.
V	1.000	.	.	.	.

## Computation

### Algorithms

The Quasi-Newton method is described in Fletcher (1972) and is sometimes called modified Fletcher/Powell. Modifications include the LDL' Cholesky factorization of the updated Hessian matrix. It is the same algorithm employed in SERIES for ARIMA estimation. The Simplex method is adapted from O'Neill (1971), with several revisions noted in Griffiths and Hill (1985).

The loss function is computed in two steps. First, the model statement is evaluated for a case using current values of the parameters and data. Second, the LOSS statement is evaluated using ESTIMATE (computed as the result of the model statement evaluation) and other parameter and data values. These two steps are repeated for all cases, over which the result of the loss function is summed. The summed LOSS is then minimized by the Quasi-Newton or Simplex procedure. Step halvings are used in the minimizations when model or loss statement evaluations overflow or result in illegal values. If repeated step halvings down to machine epsilon (error limit) fail to remedy this situation, iterations cease with an "Illegal values" message.

Asymptotic standard errors are computed by the central differencing finite approximation of the Hessian matrix. Some nonlinear regression programs compute standard errors by squaring the Jacobian matrix of first derivatives. Others use different methods altogether. For linear models, all valid methods produce identical results. For some nonlinear models, however, the results may differ. The Hessian approach, which works well for nonlinear regression, is also ideally suited for NONLIN's maximum likelihood estimation.

### Missing Data

Missing values are handled according to the conventions of SYSTAT. That is, missing values propagate in algebraic expressions. For example, "X + ." is a missing value. The expression "X = ." is not missing, however. It is 1 if X is missing and 0 if not. Thus, you can use logical expressions to put conditions on model or loss functions; consider the following loss function:

$$(X \neq .) * (Y - \text{ESTIMATE})^2 + (X = .) * (Z - \text{ESTIMATE})^2$$



Illegal expressions (such as division by 0 and negative square roots) are set to missing values. If this happens when computing the loss statement for a particular case, the loss function is set to an extremely large value ( $10^{299}$ ). This way, parameter estimates are forced to move away from regions of the parameter space that yield illegal function evaluations.

Overflows (such as a positive number with an extremely large exponent) are set to machine overflow ( $10^{299}$ ). Negative overflows are set to the negative of this value. Overflows usually cause the loss function to be large, so the program is forced to move away from estimates that produce overflows.

These features mean that NONLIN tends to “crash” less frequently than most other nonlinear estimation programs. It will continue for several iterations to try parameter values that lower the loss value, even when some of these lead to a seemingly hopeless result. It is your responsibility to check whether final estimates are reasonable, however, by using both estimation methods, different starting values, and other options.

## References

- Bates, D. M. and Watts, D. G. (1988). *Nonlinear regression and its applications*. New York: John Wiley & Sons.
- Clarke, G. P. Y. (1987). Approximate confidence limits for a parameter function in nonlinear regression, *Journal of the American Statistical Association*, 82, 221–230.
- Cook, R. D. and Weisberg, S. (1990). Confidence curves in nonlinear regression, *Journal of the American Statistical Association*, 85, 544–551.
- Cox, D. R. and Snell, E. J. (1989). *The analysis of binary data*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.
- Fletcher, R. (1972). *FORTRAN subroutines for minimization by Quasi-Newton methods*. AERE R. 7125.
- Griffiths, P. and Hill, I. D. (1985). *Applied statistics algorithms*. Chichester: Ellis Horwood Limited.
- Gilfoil, D. M. (1982). Warming up to computers: A study of cognitive and affective interaction over time. In *Proceedings: Human factors in computer systems*. Washington, D.C.: Association for Computing Machinery.
- Greco, W. R., Priore, R.L., Sharma, M., and Korytnyk, W. (1982). ROSFIT: An enzyme kinetics nonlinear regression curve fitting package for a microcomputer. *Computers and Biomedical Research*, 15, 39–45.
- \* Hill, M. A. and Engelman, L. (1992). Graphical aids for nonlinear regression and discriminant analysis. *Computational Statistics*, vol. 2, Y. Dodge and J. Whittaker, eds. Proceedings of the 10th Symposium on Computational Statistics Physica-Verlag, 111–126.
- Jennrich, R. I. and Moore, R. H. (1975). Maximum likelihood estimation by means of nonlinear least-squares. *Proceedings of the Statistical Computing Section*, American Statistical Association, 57–65.
- Kutner, M.H., Nachtsheim, C.J., Neter, J., and Li, W (2004). *Applied linear statistical models*, 5th ed. New York: McGraw-Hill/Irwin.
- \* Montgomery, D.C., Peck, E.A., and Vining, G.G. (2006). *Introduction to linear regression analysis*, 4th ed. Hoboken, N.J.: Wiley-Interscience.
- O'Neill, R. (1971). Functions minimization using a simplex procedure. Algorithms AS 47. *Applied Statistics*, 338.
- Rosenbrock, H. H. (1960). An automatic method for finding the greatest or least value of a function. *Journal of Computing*, 3, 175–184.
- \* Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: John Wiley & Sons.

(\* indicates additional references.)

# Nonparametric Tests

Leland Wilkinson

(modified by Mangalmurti Badgular and Ravindra Jore)

Nonparametric Tests perform nonparametric tests for groups of cases and pairs of variables. Tests are available for two or more independent groups of cases, two or more dependent variables, and for the distribution of a single variable.

Nonparametric tests do not assume that the data conform to a particular probability distribution. Nonparametric models are often appropriate when the usual parameters, such as mean and standard deviation based on normal theory, do not apply. Usually, however, some other assumptions about shape and continuity are made. Note that if you can find normalizing transformations for your data that allow you to use parametric tests, you will usually be better off doing so.

Several nonparametric tests are available. The Kruskal-Wallis test and the two-sample Kolmogorov-Smirnov test measure differences of a single variable across two or more independent groups of cases. The sign test, the Wilcoxon signed-rank test, the Friedman test, and the Quade test measure differences among related samples. The one-sample Kolmogorov-Smirnov test, the Anderson-Darling test, and the Wald-Wolfowitz runs test examine the distribution of a single variable.

Many nonparametric statistics are computed elsewhere in SYSTAT. Correlations calculates matrices of coefficients, such as Spearman's rho, Kendall's tau-b, Guttman's mu<sup>2</sup>, Goodman-Kruskal gamma, Goodman-Kruskal lambda and Cramer's V. Descriptive Statistics offers stem-and-leaf plots, and Box Plot offers box plots with medians and quartiles. Time Series can perform nonmetric smoothing. Crosstabs can be used for chi-square tests of independence. Multidimensional Scaling (MDS) and Cluster Analysis work with nonmetric data matrices. Finally, you can use Rank to compute a variety of rank-order statistics.

Resampling procedures are available in this feature.



**Note:** Beware of using nonparametric procedures to rescue bad data. In most cases, these procedures were designed to apply to categorical or ranked data, such as rank judgments and binary data. If you have data that violate distributional assumptions for linear models, you should consider transformations or robust models before retreating to nonparametrics.

## Statistical Background

**Nonparametric** statistics is a misnomer. The term is ordinarily used to describe a heterogeneous group of procedures that require relatively minimal assumptions about the shape of distributions underlying an analysis. Frequently, however, nonparametric models include parameters. These parameters are not necessarily ones like  $\mu$  and  $\sigma$ , which we see in typical parametric tests based on normal theory, but they are parameters in a class of mathematical functions nonetheless.

In this context, a better term for nonparametric is **distribution-free**. That is, the data for this class of statistical tests are not assumed to follow a specific probability distribution. This does not mean, however, that we make *no* assumptions about distributions in nonparametric methods. For example, in the Mann-Whitney and Kruskal-Wallis tests, we assume that the underlying populations are continuous and have the same shape.

### Rank (Ordinal) Data

An aspect of many nonparametric tests is that they are invariant under *rank-order* transformations of the data values. In other words, we may change actual data values as long as we preserve relative ranks, and the results of our hypothesis tests will not change. Data that can be replaced by rank-order values without losing information are often called **rank** or **ordinal data**. For example, if we believe that the list (-25, 54, 107.6, 3400) contains only ordinal information, then we can replace it with the list (1, 2, 3, 4) without loss of information.



## ***Categorical (Nominal) Data***

Some nonparametric methods are invariant under *permutation* transformations. That is, we can interchange data values and get the same results, provided we keep all cases with one value before transformation and single valued after transformation. Data that can be treated like this are often called **categorical** or **nominal**. For example, if we believe the list (1, 1, 5, 5, 10, 10, 10) contains only nominal information, then we can replace it with the list (red, red, green, green, blue, blue, blue) without loss of information.

## ***Robustness***

Sometimes, we may think our data contain more than nominal or ordinal information, but we want to be extremely conservative. For example, our data may contain extreme outliers. We could eliminate these outliers, downweight them, or apply some nonlinear transformation to reduce their influence. An alternative, however, would be to use a nonparametric test based on ranks. If we can afford to lose some power by using a nonparametric test, we can gain robustness. If we find significant results with a nonparametric test, no skeptic can challenge us on the basis of scale artifacts or outliers. This is not to say that you should retreat to nonparametric methods every time you find a histogram that does not look normal. If you can find a simple normalizing transformation that works, such as logging the data, you will almost always be better off using normal parametric methods. For more information about nonparametric statistical methods, see Hollander and Wolfe (1999), Lehmann and D'Abrera (1998), Mosteller and Rourke (1973), Siegel and Castellan (1988).

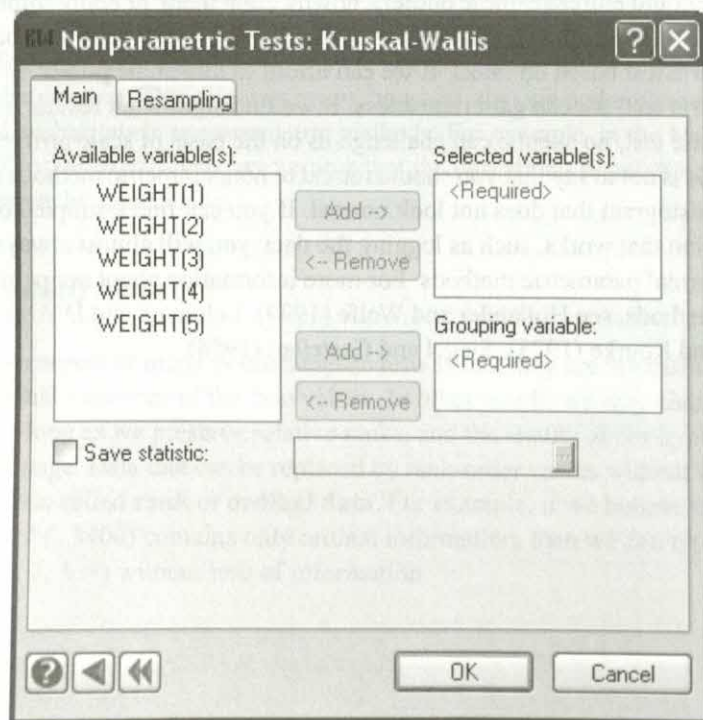
## *Nonparametric Tests for Independent Samples in SYSTAT*

### *Kruskal-Wallis Test Dialog Box*

For the Kruskal-Wallis test, the values of a variable are transformed to ranks (ignoring group membership) to test that there is no shift in the center of the groups (that is, the centers do not differ). This is the nonparametric analog of a one-way analysis of variance. When there are only two groups, this procedure reduces to the Mann-Whitney test, the nonparametric analog of the two-sample  $t$  test.

To open the Kruskal-Wallis Test dialog box, from the menus choose:

Analyze  
Nonparametric Tests  
Kruskal-Wallis...



**Selected variable(s).** SYSTAT computes a separate test for each variable in the Selected variable(s) list.

**Grouping variable.** The grouping variable can be a string or numeric.

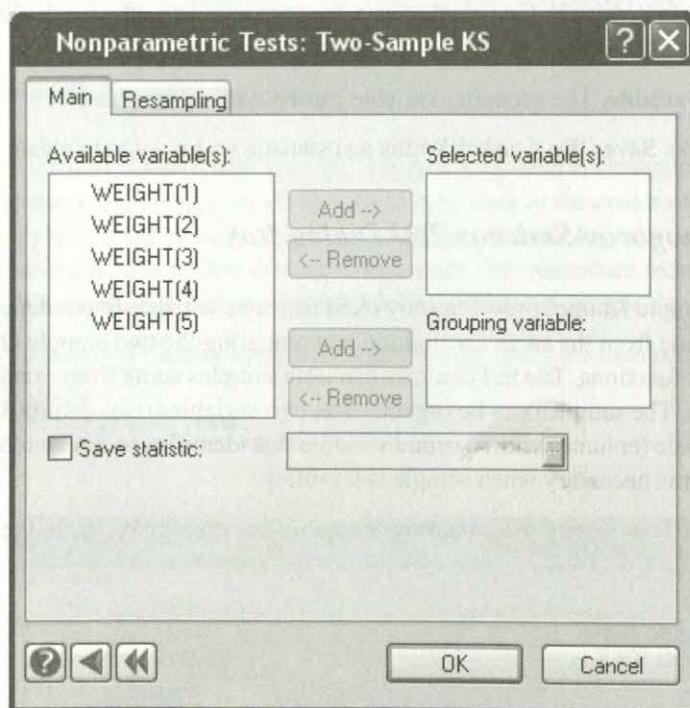
**Save statistic.** Saves the Kruskal-Wallis test statistic and  $p$ -value to a data file.

### ***Two-Sample Kolmogorov-Smirnov Test Dialog Box***

The two-sample Kolmogorov-Smirnov (KS) test tests whether two independent samples come from the same distribution by comparing the two-sample cumulative distribution functions. The test assumes that both samples come from exactly the same distribution. The samples can be organized as two variables (two columns) or as a single variable (column) with a second variable that identifies group membership. The latter layout is necessary when sample sizes differ.

To open the Two-Sample Kolmogorov-Smirnov Test dialog box, from the menus choose:

- Analyze
- Nonparametric Tests
- Two-Sample KS...



**Selected variable(s).** If each sample is a separate variable, both variables must be selected. Selecting three or more variables yields a separate test for each pair of variables. If you select only one variable, you must identify the grouping variable. If you do not select any of the variables, two sample tests are computed using numeric variables.

**Grouping variable.** If the grouping variable has three or more levels, separate tests of each pair of levels result. Selecting multiple variables and a grouping variable yields a test comparing the groups for the first variable only.

**Save statistic.** Saves the KS test statistics and  $p$ -values for all pairs of groups to a data file.



## Using Commands

First, specify your data with *USE filename*. Continue with:

```

NPAR
  SAVE or WORK filename
  KRUSKAL varlist*grpvar /SAMPLE = BOOT(m,n)
                                = JACK
                                = SIMPLE(m,n)
  KS varlist*grpvar /SAMPLE = BOOT(m,n)
                           = JACK
                           = SIMPLE(m,n)

```

## Nonparametric Tests for Related Variables in SYSTAT

A need for comparing variables frequently arises in 'before' and 'after' studies, where each subject is measured before and after a treatment. Here your goal is to determine if any difference in response can be attributed to chance alone. As a test, researchers often use the sign test or the Wilcoxon signed-rank test. For these tests, the measurements need not be collected at different points in time; they simply can be two measures on the same scale for which you want to test differences. If you have more than two measures for each subject, the Friedman test can be used.

### Sign Test Dialog Box

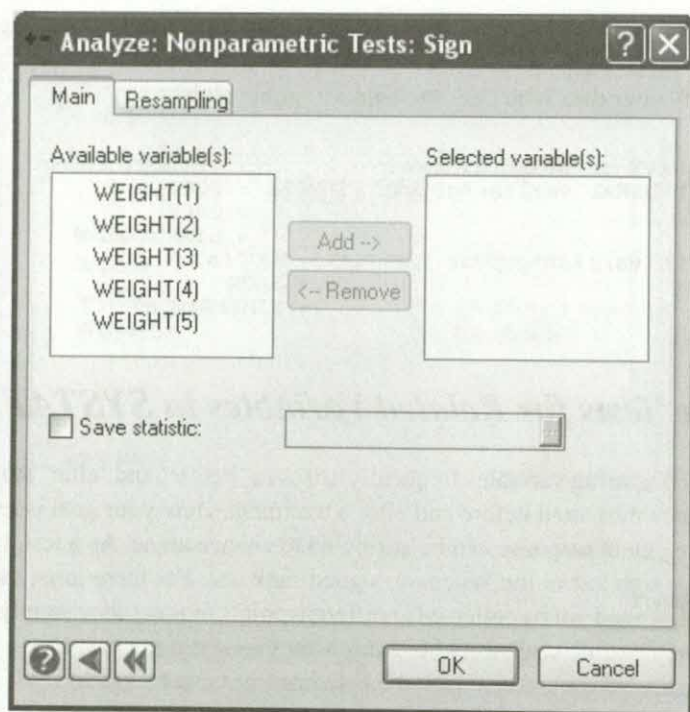
The sign test compares two related samples and is analogous to the paired *t* test. For each case, the sign test computes the sign of the difference between two variables. This test is attractive because of its simplicity and the fact that the variance of the first measure in each pair may differ from that of the second. However, you may be losing information since the magnitude of each difference is ignored.

To open the Sign Test dialog box, from the menus choose:

```

Analyze
  Nonparametric Tests
    Sign...

```



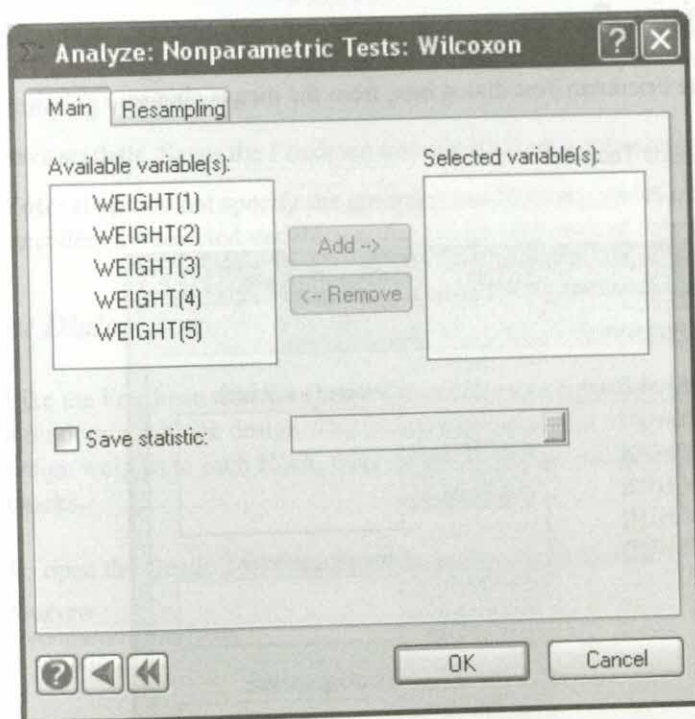
**Selected variable(s).** Selecting three or more variables yields separate tests for each pair of variables.

**Save statistic.** Saves the matrix of test statistics and the matrix of  $p$ -values to a data file.

### ***Wilcoxon Signed-Rank Test Dialog Box***

The Wilcoxon test compares the rank values of the variables you select, pair by pair, and displays the count of positive and negative differences. For ties, the average rank is assigned. It then computes the sum of ranks associated with positive differences and the sum of ranks associated with negative differences. The test statistic is the lesser of the two sums of ranks. To open the Wilcoxon Signed-Rank Test dialog box, from the menus choose:

Analyze  
Nonparametric Tests  
Wilcoxon...



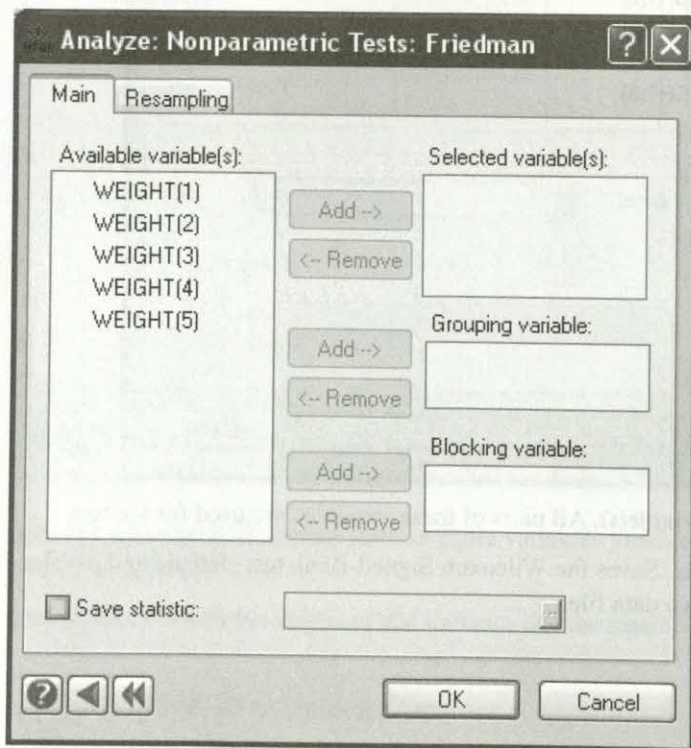
**Selected variable(s).** All pairs of these variables are used for the test.

**Save statistic.** Saves the Wilcoxon Signed-Rank test statistic and  $p$ -value for all pairs of groups to a data file.

## Friedman Test Dialog Box

To open the Friedman Test dialog box, from the menus choose:

Analyze  
Nonparametric Tests  
Friedman...



**Selected variable(s).** The Friedman test is performed separately for each of the selected variables using grouping and blocking variables, if specified.

**Grouping variable.** Select the grouping variable to define the levels of the first factor of the two-way data. The Friedman test tests the equality of the levels of the grouping effect. If you specify the grouping variable, you must specify the blocking variable also.



**Blocking variable.** Select the blocking variable to define the levels of the second factor of the two-way data. If you specify the blocking variable, you must specify the grouping variable also.

**Save statistic.** Saves the Friedman test statistic and  $p$ -value to a data file.

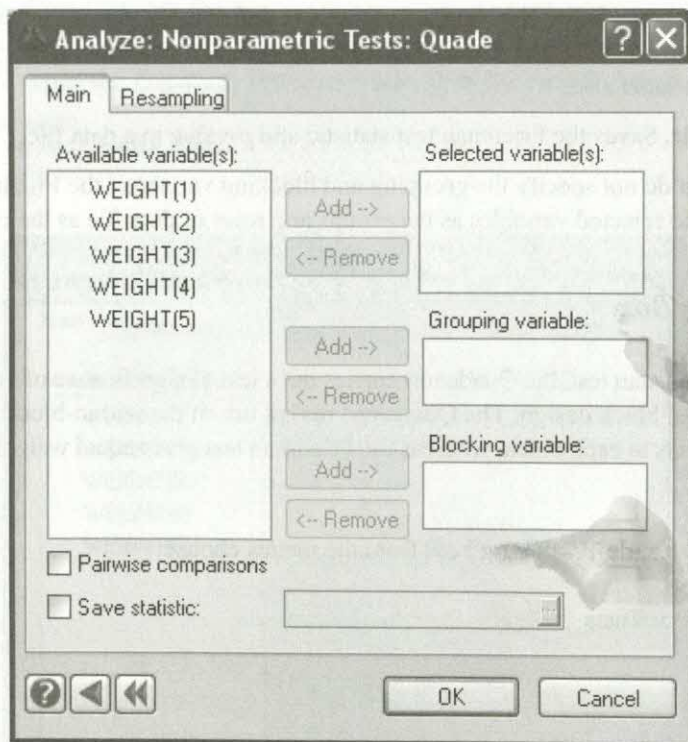
**Note:** If you do not specify the grouping and blocking variables, the Friedman test considers the selected variables as the groups and rows of data file as the blocks.

### ***Quade Test Dialog Box***

Like the Friedman test, the Quade test carries out a test of significance of one factor in a randomized block design. The Quade test makes use of the within-block range to assign weights to each block, whereas the Friedman test gives equal weights to all the blocks.

To open the Quade Test dialog box, from the menus choose:

Analyze  
Nonparametric Tests  
Quade...



**Selected variable(s).** If more than one variable is selected, Quade's analysis is carried out separately for each variable.

**Grouping variable.** Select the grouping variable to define the levels of the first factor of the two-way data. The Quade test tests the equality of the levels of the grouping effect. If you specify the grouping variable, you must specify the blocking variable also.

**Blocking variable.** Select the blocking variable to define the levels of the second factor of the two-way data. If you specify the blocking variable, you must specify the grouping variable also.

**Note:** If you do not specify the grouping and blocking variables, the Quade test considers the selected variables as the groups and rows of data file as the blocks.

**Pairwise comparisons.** Check the Pairwise comparisons option to perform the pairwise (multiple) comparisons test among different levels of the grouping variable.

**Save statistic.** Saves the Quade test statistic and  $p$ -value to a data file. If the Pairwise comparisons option is selected it saves the statistics and  $p$ -value for each pair of group levels.

## Using Commands

First, specify your data with *USE filename*. Continue with:

```

NPAR
  SAVE or WORK filename
  SIGN varlist/SAMPLE = BOOT(m,n)
                        = JACK
                        = SIMPLE(m,n)
  WILCOXON varlist/SAMPLE = BOOT(m,n)
                        = JACK
                        = SIMPLE(m,n)
  FRIEDMAN varlist=groupvar blockvar
  QUADE varlist=groupvar blockvar/MULTIPLE

```

## Nonparametric Tests for Single Samples in SYSTAT

### One-Sample Kolmogorov-Smirnov Test Dialog Box

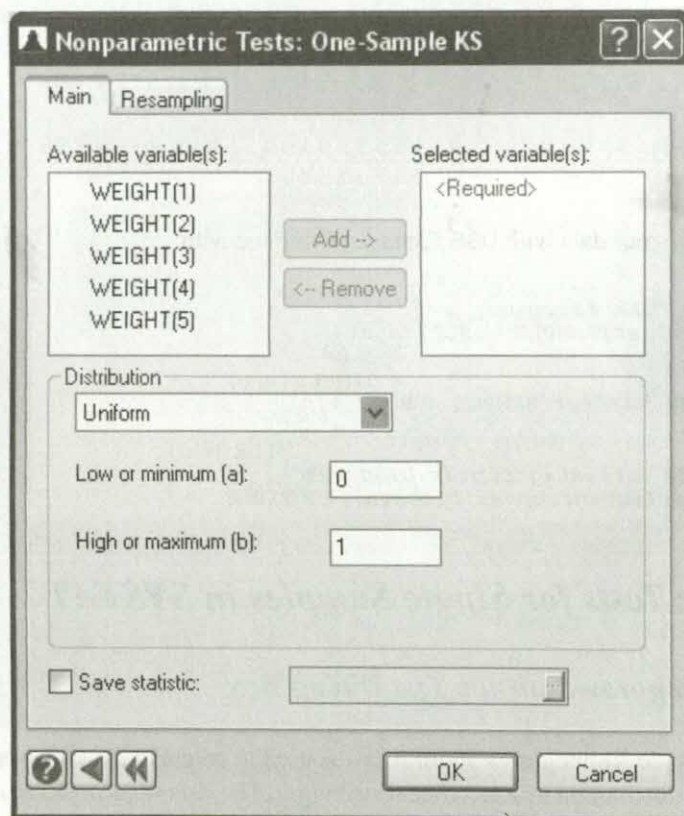
The one-sample Kolmogorov-Smirnov test is used to compare the shape and location of a sample distribution to a specified distribution. The Kolmogorov-Smirnov test and its generalizations are among the handiest of distribution-free tests. The test statistic is based on the maximum difference between two cumulative distribution functions (CDF). In the one-sample test, one of the CDF's is continuous and the other is discrete. Thus, it is a companion test to a probability plot.

To open the One-Sample Kolmogorov-Smirnov Test dialog box, from the menus choose:

```

Analyze
  Nonparametric Tests
    One-Sample KS...

```



**Selected variable(s).** The One-Sample Kolmogorov-Smirnov test is performed separately for each of the variables in the selected list.

**Distribution.** Allows you to choose the test distribution. Many options allow you to specify parameters of the hypothesized distribution. For example, if you choose a Uniform distribution, you can specify values for min and max. Distributions include:

- **Benford's Law.** Compares the data to the Benford's law( $B$ ) distribution.
- **Binomial.** Compares the data to the binomial ( $n, p$ ) distribution.
- **Discrete uniform.** Compares the data to the discrete uniform( $N$ ) distribution.
- **Geometric.** Compares the data to the geometric ( $p$ ) distribution.
- **Hypergeometric.** Compares the data to the hypergeometric ( $N, m, n$ ) distribution.
- **Logarithmic series.** Compares the data to the logarithmic series ( $\theta$ ) distribution.



- **Negative binomial.** Compares the data to the negative binomial ( $k, p$ ) distribution.
- **Poisson.** Compares the data to the Poisson ( $\lambda$ ) distribution.
- **Zipf.** Compares the data to the Zipf( $\text{shp}$ ) distribution.
- **Beta.** Compares the data to the beta( $\text{shp1}, \text{shp2}$ ) distribution.
- **Cauchy.** Compares the data to the Cauchy( $\text{loc}, \text{sc}$ ) distribution.
- **Chi-square.** Compares the data to the chi-square( $\text{df}$ ) distribution.
- **Double exponential (Laplace).** Compares the data to the Laplace ( $\text{loc}, \text{sc}$ ) distribution.
- **Erlang.** Compares the data to the Erlang( $\text{shp}, \text{sc}$ ) distribution.
- **Exponential.** Compares the data to the exponential( $\text{loc}, \text{sc}$ ) distribution.
- **F.** Compares the data to the F( $\text{df1}, \text{df2}$ ) distribution.
- **Gamma.** Compares the data to the gamma ( $\text{shp}, \text{sc}$ ) distribution.
- **Gompertz.** Compares the data to the Gompertz ( $b, c$ ) distribution.
- **Gumbel.** Compares the data to the Gumbel ( $\text{loc}, \text{sc}$ ) distribution.
- **Inverse Gaussian (Wald).** Compares the data to the Wald ( $\text{loc}, \text{sc}$ ) distribution.
- **Logistic.** Compares the data to the logistic ( $\text{loc}, \text{sc}$ ) distribution.
- **Loglogistic.** Compares the data to the loglogistic ( $\text{logsc}, \text{shp}$ ) distribution.
- **Lognormal.** Compares the data to the lognormal ( $\text{loc}, \text{sc}$ ) distribution.
- **Logitnormal.** Compares the data to the logit normal ( $\text{loc}, \text{sc}$ ) distribution.
- **Non-central chi-square.** Compares the data to the non-central chi-square( $\text{df}, \text{delta}$ ) distribution.
- **Non-central F.** Compares the data to the non-central F( $\text{df1}, \text{df2}, \text{delta}$ ) distribution.
- **Non-central t.** Compares the data to the non-central t( $\text{df}, \text{delta}$ ) distribution.
- **Normal.** Compares the data to the normal ( $\text{loc}, \text{sc}$ ) distribution.
- **Pareto.** Compares the data to the Pareto ( $\text{thr}, \text{shp}$ ) distribution.
- **Rayleigh.** Compares the data to the Rayleigh( $\text{sc}$ ) distribution.
- **Smallest extreme value.** Compares the data to the smallest extreme value ( $\text{loc}, \text{sc}$ ) distribution.
- **Studentized maximum modulus.** Compares data to the studentized maximum modulus ( $k, \nu$ ) distribution.
- **Studentized range.** Compares the data to the Studentized range ( $k, \text{df}$ ) distribution.
- **t.** Compares the data to the t( $\text{df}$ ) distribution.

- **Triangular.** Compares the data to the triangular( $a$ ,  $b$ ,  $c$ ) distribution.
- **Uniform.** Compares the data to the uniform(min, max) distribution.
- **Weibull.** Compares the data to the Weibull(sc, shp) distribution.
- **Lilliefors.** The Lilliefors test uses the standard normal distribution. The variables you select are automatically standardized, and the test determines whether the standardized versions are normally distributed.

**Note:** Lilliefors is not a distribution but is included under 'distributions' for convenience. It can be used to test normality when the parameters are not specified.

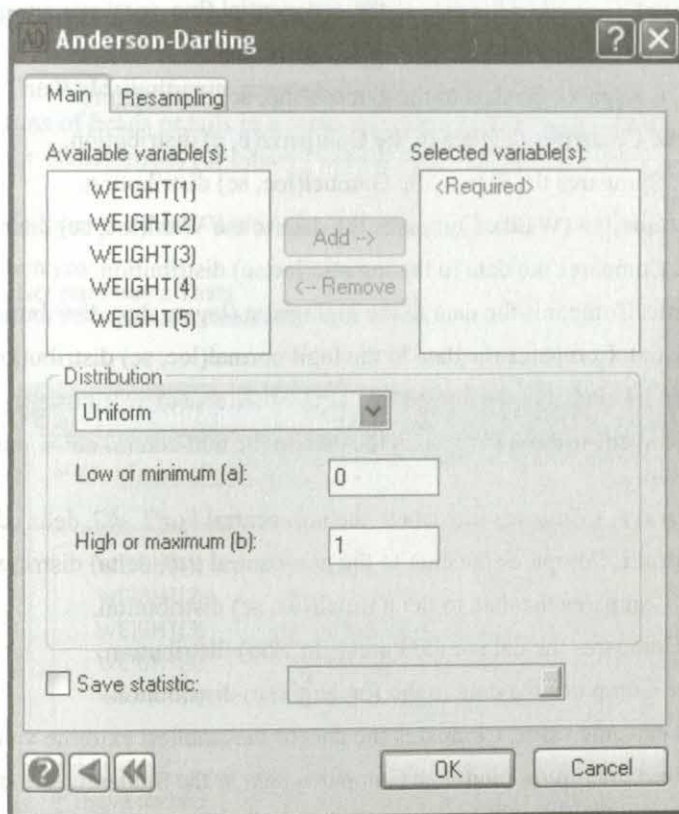
**Save statistic.** Saves the KS test statistic and  $p$ -value to a data file.

### ***Anderson-Darling Test Dialog Box***

The Anderson-Darling test (Anderson and Darling, 1952, 1954) is a standard goodness-of-fit test. It is based on the squared difference between the theoretical and empirical distribution functions, weighted by  $[F(x)(1-F(x))]^{-1}$ . This test has good power properties over a wide range of alternative distributions.

To open the Anderson-Darling Test dialog box, from the menus choose:

Analyze  
Nonparametric Tests  
Anderson-Darling...



**Selected variable(s).** The Anderson-Darling test is performed separately for each of the variables in the selected list.

**Distribution.** Allows you to choose the test distribution. Many options allow you to specify parameters of the hypothesized distribution. For example, if you choose a Uniform distribution, you can specify values for min and max. Distributions include:

- **Beta.** Compares the data to the beta(shp1, shp2) distribution.
- **Cauchy.** Compares the data to the Cauchy(loc, sc) distribution.
- **Chi-square.** Compares the data to the chi-square(df) distribution.
- **Double Exponential(Laplace).** Compares the data to the Laplace (loc, sc) distribution.
- **Erlang.** Compares the data to the Erlang(shp, sc) distribution.



- **Exponential.** Compares the data to the exponential (loc, sc) distribution.
- **F.** Compares the data to the  $F(df1, df2)$  distribution.
- **Gamma.** Compares the data to the gamma(shp, sc) distribution.
- **Gompertz.** Compares the data to the Gompertz(b, c) distribution.
- **Gumbel.** Compares the data to the Gumbel(loc, sc) distribution.
- **Inverse Gaussian (Wald).** Compares the data to the Wald(loc, sc) distribution.
- **Logistic.** Compares the data to the logistic(loc,sc) distribution.
- **Loglogistic.** Compares the data to the loglogistic(logsc, shp) distribution.
- **Logitnormal.** Compares the data to the logit normal(loc, sc) distribution.
- **Lognormal.** Compares the data to the lognormal(loc, sc) distribution.
- **Non-central chi-square.** Compares the data to the non-central chi-square(df,delta) distribution.
- **Non-central F.** Compares the data to the non-central  $F(df1, df2, delta)$  distribution.
- **Non-central t.** Compares the data to the non-central  $t(df, delta)$  distribution.
- **Normal.** Compares the data to the normal(loc, sc) distribution.
- **Pareto.** Compares the data to the Pareto(thr, shp) distribution.
- **Rayleigh.** Compares the data to the Rayleigh(sc) distribution.
- **Smallest extreme value.** Compares the data to the smallest extreme value (loc,sc).
- **Studentized maximum modulus.** Compares data to the Studentized maximum modulus ( $k, \nu$ ) distribution.
- **Studentized range.** Compares the data to the Studentized range( $k, df$ ) distribution.
- **t.** Compares the data to the  $t(df)$  distribution.
- **Triangular.** Compares the data to the triangular( $a, b, c$ ) distribution.
- **Uniform.** Compares the data to the uniform(min, max) distribution.
- **Weibull.** Compares the data to the Weibull(sc, shp) distribution.

**Save statistic.** Saves the Anderson-Darling test statistic and  $p$ -value to a data file.

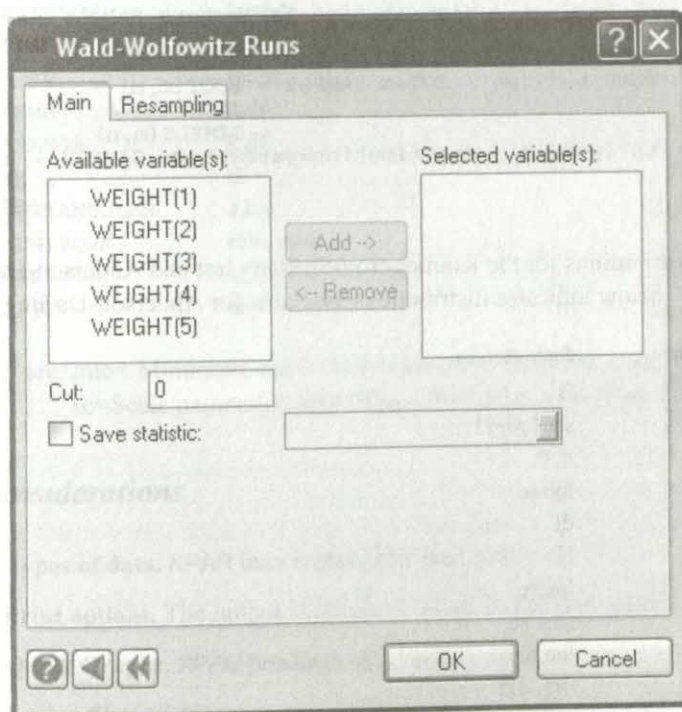


### Wald-Wolfowitz Runs Test Dialog Box

The Wald-Wolfowitz runs test detects serial patterns in a run of numbers (for example, runs of heads or tails in a series of coin tosses). The runs test measures such behavior for dichotomous (or binary) variables.

To open the Wald-Wolfowitz Runs Test dialog box, from the menus choose:

Analyze  
Nonparametric Tests  
Wald-Wolfowitz Runs...



**Selected variable(s).** Runs are calculated separately for each of the variables selected into the selected variables text box.

**Cut.** Specify a cut point value for continuous variables to determine whether values fluctuate in patterns above and below this cutpoint. This feature is useful for studying trends in residuals from a regression analysis.

**Save statistic.** Saves the number of runs, test statistic and *p*-value to a data file.

## Using Commands

First, specify your data with `USE filename`. Continue with:

```

NPAR
  SAVE or WORK filename
  AD varlist / distribution=parameters SAMPLE = BOOT(m,n)
                                           = JACK
                                           = SIMPLE(m,n)

  RUNS varlist / CUT=n SAMPLE = BOOT(m,n)
                                           = JACK
                                           = SIMPLE(m,n)

  KS varlist / distribution=parameters SAMPLE = BOOT(m,n)
                                           = JACK
                                           = SIMPLE(m,n)

```

Possible distributions for the Kolmogorov-Smirnov test and Anderson-Darling test include ('\*' below indicates distributions available for Anderson-Darling test):

Distribution	Parameters
BENFORD	B
*BETA	shp1,shp2
BINOMIAL	n, p
*CAUCHY	loc,sc
*CHISQ	df
DUNIFORM	N
*DEXP	loc,sc
ERLANG	shp, sc
*EXP	loc,sc
*F	df1, df2
*GAMMA	shp,sc
GEOMETRIC	p
*GOMPERTZ	b,c
*GUMBEL	loc,sc
HGEOMETRIC	N,m,n
*IGAUSSIAN	loc,sc
LILLIEFORS	

L SERIES	theta
*LOGISTIC	loc,sc
*ENORMAL	loc,sc
*LLOGISTIC	logsc, shp
*LNORMAL	loc,sc
NBINOMIAL	k,p
*NCCHISQ	df, delta
*NCF	df1, df2, delta
*NCT	df, delta
*NORMAL	loc,sc
*PARETO	thr,shp
POISSON	lambda
*RAYLEIGH	sc
*SEV	loc,sc
*SMM	k,df
*RANGE	k,df
*t	df
*TRIANGULAR	a,b,c
*UNIFORM	min, max
*WEIBULL	sc,shp
ZIPF	shp

**Note:** min= Minimum; max= Maximum; loc=Location parameter;  
 sc=Scale parameter; shp=Shape parameter; thr= Threshold parameter

## Usage Considerations

**Types of data.** NPAR uses rectangular data only.

**Print options.** The output is standard for all PLENGTH options.

**Quick Graphs.** NPAR produces no Quick Graphs.

**Saving files.** NPAR saves test statistics and *p*-values into a SYSTAT data file.

**BY groups.** You can perform tests using a BY variable. The output includes separate tests for each level of the BY variable.

**Case frequencies.** NPAR uses a FREQUENCY variable (if present) to increase the number of cases in the analysis.

Case weights. WEIGHT is not available in NPAR.

## Examples

### Example 1

#### Kruskal-Wallis Test

For two or more independent groups, the Kruskal-Wallis test statistic tests whether the  $k$  samples come from identically distributed populations. If the grouping variable has only two levels, the Mann-Whitney (Wilcoxon) statistic is reported. For two groups, the Kruskal-Wallis test and the Mann-Whitney  $U$  statistic are analogous to the independent groups  $t$  test.

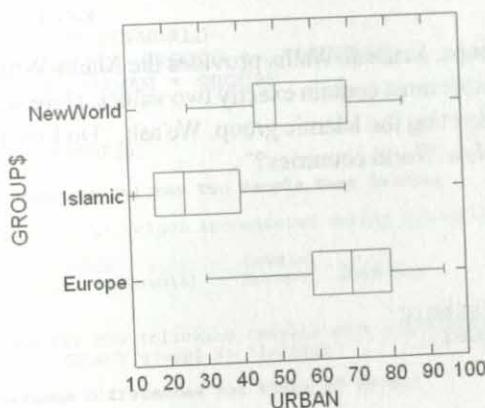
In this example, we compare the percentage of people who live in cities (*URBAN*) for three groups of countries: European, Islamic, and New World. We use the *OURWORLD* data file that has one record for each of the 57 countries with the variables *URBAN* and *GROUP\$*. We include a box plot of *URBAN* grouped by *GROUP\$* to illustrate the test.

The input is:

```
NPAR
USE OURWORLD
DENSITY URBAN * GROUP$ / BOX TRANS
KRUSKAL URBAN * GROUP$
```



The output is:



Kruskal-Wallis One-way Analysis of Variance for 57 Cases

Categorical Values Encountered during Processing are

Variables	Levels
GROUP\$ (3levels)	Europe Islamic NewWorld

Dependent Variable	URBAN
Grouping Variable	GROUP\$

Group	Count	Rank Sum
Europe	19	765.000
Islamic	16	198.000
NewWorld	21	633.000

Kruskal-Wallis Test Statistic : 25.759

p-value is 0.000 assuming Chi-square Distribution with 2 df

In the box plot, the median of each distribution is marked by the vertical bar inside the box: the median for European countries is 69%; for Islamic countries, 24.5%; and for New World countries, 50%. We ask, "Is there a difference in typical values of *URBAN* among these groups of countries?"

Looking at the Kruskal-Wallis results, we find a  $p$ -value  $< 0.0005$ . We conclude that urbanization differs markedly across the three groups of countries.

## Example 2

### Mann-Whitney Test

When there are only two groups, Kruskal-Wallis provides the Mann-Whitney test. Note that your grouping variable must contain exactly two values. Here we modify the Kruskal-Wallis example by deleting the Islamic group. We ask, "Do European nations tend to be more urban than New World countries?"

The input is:

```
NPAP
USE OURWORLD
SELECT GROUP$ <> 'ISLAMIC'
KRUSKAL URBAN * GROUP$
```

The output is:

#### Kruskal-Wallis One-way Analysis of Variance for 57 Cases

Categorical Values Encountered during Processing are

Variables	Levels
GROUP\$ (2levels)	Europe NewWorld

Data for the following results were selected according to  
SELECT group\$ <> 'Islamic'

Dependent Variable	URBAN
Grouping Variable	GROUP\$

Group	Count	Rank Sum
Europe	19	475.000
NewWorld	21	345.000

Mann-Whitney U Test Statistic	: 285.000
p-value	: 0.020
Chi-square Approximation	: 5.370
df	: 1

The percentage of the population living in urban areas is significantly greater for European countries than for New World countries ( $p$ -value = 0.02).

### Two-Sample Kolmogorov-Smirnov Test

The two-Sample Kolmogorov-Smirnov test measures the discrepancy between two-sample cumulative distribution functions.

In this example, we test if the distributions of *URBAN*, the proportion of people living in cities, for European and New World countries have the same mean, standard deviation, and shape.

The input is:

```

NPAR
USE OURWORLD
SELECT GROUP$ <> 'ISLAMIC'
KS URBAN * GROUP$

```

The output is:

#### Kolmogorov-Smirnov Two Sample Test Results

Categorical Values Encountered during Processing are

Variables	Levels
GROUP\$ (2 levels)	Europe NewWorld

Data for the following results were selected according to  
 SELECT group\$ <> 'Islamic'

#### Maximum Differences for Pairs of Groups

	Europe	NewWorld
Europe	0.000	
NewWorld	0.519	0.000

#### Two-Sided Probabilities

	Europe	NewWorld
Europe	1.000	
NewWorld	0.009	1.000

From the  $p$ -value, we can conclude that the population distributions for European and New World countries are different.

### Example 3

#### Sign Test

Here, for a sample of countries (not subjects), we ask, "Does life expectancy differ for males and females?" Using the *OURWORLD* data, we compare *LIFEEXPF* and *LIFEEXPM*, using stem-and-leaf plots to illustrate the distributions. The sign test counts the number of times male life expectancy is greater than that for females and vice versa.

The input is:

```

USE OURWORLD
STEM LIFEEXPF LIFEEXPM / LINES=10
NPAR
SIGN LIFEEXPF LIFEEXPM

```

The output is:

**Stem and Leaf Plot of Variable: LIFEEXPF, N = 57**

Minimum : 44.000  
Lower Hinge : 65.000  
Median : 75.000  
Upper Hinge : 79.000  
Maximum : 83.000

```

 4  4
 4  679
 5  0234
 5  55667
 6  4
 6 H 567788889
 7  01344
 7 M 5666777778889999
 8  0000111111223

```

**Stem and Leaf Plot of Variable: LIFEEXPM, N = 57**

Minimum : 40.000  
Lower Hinge : 61.000  
Median : 68.000  
Upper Hinge : 73.000  
Maximum : 75.000

```

 4  0
*** Outside Values ***
 4  56789
 5  122334
 5  6
 6 H 01222444
 6 M 5556778899
 7 H 00111122333333334444
 7  55555

```

**Sign Test Results**

**Counts of Differences (Row Variable Greater than Column)**

	LIFEEXPM	LIFEEXPF
LIFEEXPM	0.000	2.000
LIFEEXPF	55.000	0.000

**Two-Sided Probabilities for Each Pair of Variables**

	LIFEEXPM	LIFEEXPF
LIFEEXPM	1.000	
LIFEEXPF	0.000	1.000

For each case, SYSTAT first reports the number of differences that were positive and the number that were negative. In two countries (Afghanistan and Bangladesh), the males live longer than the females; the reverse is true for the other 55 countries. Note that the layout of this output allows reports for many pairs of variables.

In the two-sided probabilities panel, the smaller count of differences (positive or negative) is compared to the total number of nonzero differences. SYSTAT computes a sign test on all possible pairs of specified variables. For each pair, the difference



between values on each case is calculated, and the number of positive and negative differences is printed. The lesser of the two types of differences (positive or negative) is then compared to the total number of nonzero differences. From this comparison, the probability is computed according to the binomial (for a total less than or equal to 25) or a normal approximation to the binomial (for a total greater than 25). A correction for continuity (0.5) is added to the normal approximation's numerator, and the denominator is computed from the null value of 0.5. The large sample test is thus equivalent to a chi-square test for an underlying proportion of 0.5. The probability for our test is 0.000 (or  $< 0.0005$ ). We conclude that there is a significant difference in life expectancy; females tend to live longer.

#### Example 4 Wilcoxon Test

Here, as in the sign test example, we ask, "Does life expectancy differ for males and females?"

The input is:

```
USE OURWORLD
NPAR
WILCOXON LIFEEXPM LIFEEXPF
```

The output is:

##### Wilcoxon Signed Ranks Test Results

Counts of Differences (Row Variable Greater than Column)

	LIFEEXPM	LIFEEXPF
LIFEEXPM	0.000	2.000
LIFEEXPF	55.000	0.000

$Z = (\text{Sum of Signed ranks}) / \text{Square root}(\text{Sum of Squared ranks})$

	LIFEEXPM	LIFEEXPF
LIFEEXPM	0.000	
LIFEEXPF	6.535	0.000

##### Two-Sided Probabilities using Normal Approximation

	LIFEEXPM	LIFEEXPF
LIFEEXPM	1.000	
LIFEEXPF	0.000	1.000

Two-sided probabilities are computed from an approximate normal variate ( $Z$  in the output) constructed from the lesser of the sum of the positive ranks and the sum of the

negative ranks (for example, Marascuilo and McSweeney, 1977, p. 338). The  $Z$  for our test is 6.535 with a probability less than 0.0005. As with the sign test, we conclude that females tend to live longer.

### Example 5

#### Sign and Wilcoxon Tests for Multiple Variables

SYSTAT can compute a sign or Wilcoxon test on all pairs of specified variables (or all numeric variables in your file). To illustrate the layout of the output, we add two more variables to our request for a sign test: the birth-to-death ratios in 1982 and 1990.

The input is:

```

NPAR
USE OURWORLD
SIGN B_TO_D82 B_TO_D LIFEEXPM LIFEEXPF

```

The output is:

##### Sign Test Results

Counts of Differences (Row Variable Greater than Column)

	B_TO_D82	LIFEEXPM	LIFEEXPF	B_TO_D
B_TO_D82	0.000	0.000	0.000	17.000
LIFEEXPM	57.000	0.000	2.000	57.000
LIFEEXPF	57.000	55.000	0.000	57.000
B_TO_D	36.000	0.000	0.000	0.000

##### Two-Sided Probabilities for Each Pair of Variables

	B_TO_D82	LIFEEXPM	LIFEEXPF	B_TO_D
B_TO_D82	1.000			
LIFEEXPM	0.000	1.000		
LIFEEXPF	0.000	0.000	1.000	
B_TO_D	0.013	0.000	0.000	1.000

The results contain some meaningless data. SYSTAT has ordered the variables as they appear in the data file. When you specify more than two variables, there may be just a few numbers of interest. In the first column, the birth-to-death ratio in 1982 is compared with the birth-to-death ratio in 1990—and with male and female life expectancy! Only the last entry is relevant—36 countries have larger ratios in 1990 than they did in 1982. In the last column, you see that 17 countries have smaller ratios in 1990. The life expectancy comparisons you saw in the last example are in the middle of this table. In the two-sided probabilities panel, the probability for the birth-to-death ratio comparison (0.013) is at the bottom of the first column. We conclude that the ratio

is significantly larger in 1990 than it was in 1982. Does this mean that the number of births is increasing or that the number of deaths is decreasing?

### Example 6 Friedman Test

The following example is from Kutner, Nachtsheim, Neter and Li (2004). Five blocks of judges were given the task of analyzing three treatments. We are interested in testing the equality of the treatments. These data are in the file *BLOCK*.

The input is:

```
NPART
USE BLOCK
FRIEDMAN JUDGMENT=TREAT BLOCK
```

The output is:

#### Friedman Two-Way Analysis of Variance Results for 15 Cases

Categorical Values Encountered during Processing are

Variables	Levels
TREAT (3levels)	1.000 2.000 3.000
BLOCK (5levels)	1.000 2.000 3.000 4.000 5.000

Dependent Variable	JUDGMENT
Grouping Variable	TREAT
Blocking Variable	BLOCK
Number of Groups	3
Number of Blocks	5

TREAT	Rank Sum
1	5.000
2	10.000
3	15.000

Friedman Test Statistic : 10.000  
Kendall Coefficient of Concordance : 1.000

p-value is 0.007 assuming Chi-square Distribution with 2 df

Friedman's test rejects the hypothesis at the 5% level.

**Example 7****Friedman Test for the Case with Ties**

In this example, we study the number of books sold in a week in 12 bookstores of four booksellers and ask the question: "Is there a differential preference for the books in the stores?" Friedman's test depends only on the ranks of the books in each shop and notice that there are ties in the data set. The computation for the tied case is somewhat different and SYSTAT performs this computation. The data are fictitious, but made to correspond to Example 1 in Conover (1999, pp 371-373).

The input is:

```

NPAR
USE BOOKPREF
FRIEDMAN BOOKS = BOOKSELLER STORE

```

The output is:

**Friedman Two-Way Analysis of Variance Results for 48 Cases**

Categorical Values Encountered during Processing are

Variables	Levels
BOOKSELLER (4levels)	1 2 3 4
STORE (12levels)	1 2 3 4 5
	6 7 8 9 10
	11 12

Dependent Variable	BOOKS
Grouping Variable	BOOKSELLER
Blocking Variable	STORE
Number of Groups	4
Number of Blocks	12

BOOKSELLER	Rank Sum
1	38.000
2	23.500
3	24.500
4	34.000

Friedman Test Statistic : 8.097  
 Kendall Coefficient of Concordance : 0.225

p-value is 0.044 assuming Chi-square Distribution with 3 df

Friedman's test in this case rejects the hypothesis at the 5% level.

You may note that while computing the test statistic, SYSTAT has taken note of the ties in the data. When there is a tie, the tied observations receive the same rank, which is the average of the ranks they would get in the situation with no ties. The subsequent observations get ranks that they would have got had there been no ties. Thus the sum of the ranks remains the same whether there are ties or no ties.



### Example 8

#### Quade Test for Cases with Ties

The data were collected in a survey conducted in 7 hospitals of a certain city over a 12-month period divided into 4 seasons, and the numbers of newborn babies in each season were obtained. The data set is taken from Conover (1999). The question of interest is whether the seasonal factor has any influence on the number of births.

The input is:

```

NPAR
USE BIRTHS2
QUADE BIRTHS = SEASON$ HOSPITAL$

```

The output is:

#### Quade Two-Way Analysis of Variance Results for 28 Cases

```

Dependent Variable : BIRTHS
Grouping Variable  : SEASON$
Blocking Variable   : HOSPITAL$
Number of Groups    : 4
Number of Blocks    : 7

```

Categorical Values Encountered during Processing are

Variables	Levels
SEASON\$ (4levels)	FALL SPRING SUMMER WINTER
HOSPITAL\$ (7levels)	A B C D E
	F G
SEASON\$	Weighted
	Midranks Sum
FALL	-23.000
SPRING	37.500
SUMMER	-5.250
WINTER	-9.250

Quade Test Statistic : 4.431  
 p-value is 0.017 approximated by F(3, 18) Distribution.

The Quade test rejects the null hypothesis at 5% level.

### Example 9

#### Quade Test for Multiple Comparisons

We continue with the previous example. For the *BIRTHS2* data, the Quade test rejects the null hypothesis at 5% level. We therefore need to perform a multiple comparisons test to see which pairs of means differ significantly.

The input is:

```

NPAR
USE BIRTHS2
QUADE BIRTHS = SEASON$ HOSPITAL$ / MULTIPLE

```

The output is:

#### Quade Multiple Comparisons Test for 28 Cases

Dependent Variable	BIRTHS
Grouping Variable	SEASON\$
Blocking Variable	HOSPITAL\$
Number of Groups	4
Number of Blocks	7

Categorical Values Encountered during Processing are

Variables	Levels				
SEASON\$ (4levels)	FALL	SPRING	SUMMER	WINTER	
HOSPITAL\$ (7levels)	A	B	C	D	E
	F	G			

#### Matrix of Pairwise Differences (of Weighted Midranks)

SEASON\$	FALL	SPRING	SUMMER	WINTER
FALL	0.000			
SPRING	60.500	0.000		
SUMMER	17.750	-42.750	0.000	
WINTER	13.750	-46.750	-4.000	0.000

#### Matrix of p-values

SEASON\$	FALL	SPRING	SUMMER	WINTER
FALL	1.000			
SPRING	0.003	1.000		
SUMMER	0.325	0.026	1.000	
WINTER	0.444	0.016	0.822	1.000

Thus the 4 seasons can be divided into 2 groups, one comprising *WINTER*, *SUMMER* and *FALL*, and the other *SPRING*.

### Example 10

#### One-Sample Kolmogorov-Smirnov Test for Normal Distribution

In this example, we use SYSTAT's random number generator to make a normally distributed random number and then test it for normality. We use the variable *Z* as our normal random number and the variable *ZS* as a standardized copy of *Z*. This may seem strange because normal random numbers are expected to have a mean of 0 and a standard deviation of 1. This is not exactly true in a sample, however, so we standardize the observed values to make a variable that has exactly a mean of 0 and a standard deviation of 1.

The input is:

```

RANDSAMP
UNIVARIATE ZRN(0,1) / SIZE = 50 NSAMP = 1 RSEED = 16
LET Z = S1
LET ZS = Z
STANDARDIZE ZS / SD
CSTATISTICS
DSAVE NORMAL
USE NORMAL
NPAR
KS Z ZS / NORMAL = 0,1

```

We use CSTATISTICS to examine the mean and standard deviation of our two variables. Remember, if you correlated these two variables, the Pearson correlation would be 1. Only their mean and standard deviations differ. Finally, we test Z for normality.

The output is:

	S1	Z	ZS
N of Cases	50.000	50.000	50.000
Minimum	-2.118	-2.118	-2.266
Maximum	2.103	2.103	2.036
Arithmetic Mean	0.105	0.105	0.000
Standard Deviation	0.981	0.981	1.000

**Kolmogorov-Smirnov One Sample Test using Normal(0.000, 1.000) Distribution**

Variable	N of Cases	Maximum Difference	p-value(2-tail)
Z	50	0.150	0.210
ZS	50	0.112	0.553

Why are the probabilities different? The one-sample Kolmogorov-Smirnov test pays attention to the shape, location, and scale of the sample distribution. Z and ZS have the same shape in the population (they are both normal). Because ZS has been standardized, however, it has a different location.

Thus, you should never use the Kolmogorov-Smirnov test with the normal distribution on a variable you have standardized. The probability printed for ZS is misleading. If you select Chi-Square, Normal or Uniform, you are assuming that the variable you are testing has been randomly sampled from a chi-square (with stated degrees of freedom), standard normal or uniform (0,1) population.

### ***Lilliefors Test***

Here we perform a Lilliefors test using the data generated for the one-sample Kolmogorov-Smirnov example. Note that Lilliefors automatically standardizes the variables you list and tests whether the standardized versions are normally distributed.

The input is:

```
USE NORMAL
NPAR
KS Z ZS / LILLIEFORS
```

The output is:

Kolmogorov-Smirnov One Sample Test using Normal(0.000, 1.000) Distribution

Variable	N of Cases	Maximum Difference	Lilliefors Probability (2 tail)
Z	50	0.112	0.113
ZS	50	0.112	0.113

Notice that the probabilities are smaller this time even though the Maximum Difference is the same as before. The probability values for Z and ZS are the same because this test pays attention only to the shape of the distribution and not to the location or scale. Neither significantly differs from normal.

This example was constructed to contrast Normal and Lilliefors. Many statistical package users do a Kolmogorov-Smirnov test for normality on their standardized data without realizing that they should instead do a Lilliefors test.

One last point: The Lilliefors test can be used for residual analysis in regression. Just standardize your residuals and use Nonparametric Tests to test them for normality. If you do this, you should always look at the corresponding normal probability plot.

### ***Example 11***

#### ***One-Sample Kolmogorov-Smirnov Test for Non-Central Chi-square Distribution***

Suppose a researcher wants to test if the following observations are realizations from the non-central chi-square distribution with parameters ( $df$ ,  $\delta$ ) as (1, 3.5). 0.01, 0.61, 0.30, 3.06, 0.02, 0.87, 6.50, 3.28, 0.14, 0.19, 0.39, 2.41, 1.49, 1.02, 1.67. Input this data in a column and name it as X. We will use one-sample Kolmogorov-Smirnov test.



The input is:

```

NPAR
KS X / NCCHISQ=1, 3.5

```

The output is:

Kolmogorov-Smirnov One Sample Test using Non-central Chi-square(1.00, 3.50) Distribution

Variable	N of Cases	Maximum Difference	p-value(2-tail)
X	15	0.457	0.002

From the  $p$ -value, the researcher can easily conclude that the data differ significantly from the non-central chi-square distribution with the parameters specified.

### Example 12 Anderson-Darling Test

An electrical engineer wants to test whether the life of a certain equipment is exponentially distributed with a mean life of one year. From the testing department he collects lifetimes of 20 units of that equipment as: 0.98, 2.12, 3.65, 0.65, 0.33, 0.64, 1.02, 0.25, 0.40, 1.04, 2.12, 0.58, 1.21, 0.71, 0.17, 0.14, 0.55, 0.54, 2.06, and 0.63. He then plots the data on an exponential probability paper, but is not completely satisfied with the visual inspection of the probability plot. He therefore uses the Anderson-Darling test.

The input is:

```

USE LIFE
NPAR
AD LIFE / EXP = 0,1

```

The output is:

Anderson-Darling Test using Exponential(0.00, 1.00) Distribution

Variable	N of Cases	AD Statistic	p-value
LIFE	20.000	0.701	0.556

The Anderson-Darling test indicates that equipment data can indeed be modeled as an exponential distribution with a mean life of one year.

### Example 13

#### Wald-Wolfowitz Runs Test

We use the *OURWORLD* file and cut *MIL* (dollars per person each country spends on the military) at its median and see whether countries with higher military expenditures are grouped together in the file. (Be careful when you use a cutpoint on a continuous variable, however. Your conclusions can change depending on the cutpoint you use.) We include a scatterplot of the military expenditures against the case number (order of each country in the file), adding a dotted line at the cutpoint of 53.889.

The input is:

```

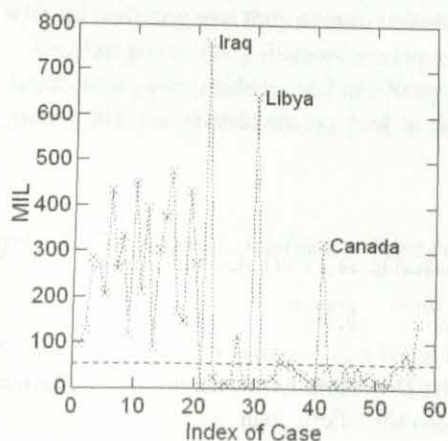
NPAR
USE OURWORLD
RUNS MIL / CUT=53.889
IF (COUNTRY$='Iraq' or COUNTRY$='Libya' or COUNTRY$='Canada'),
    THEN LET COUNTRY2$=COUNTRY$
PLOT MIL / LINE DASH=11 YLIM=53.9 LABEL=COUNTRY2$ SYMBOL=2,
    CSIZE=2

```

The output is:

Wald-Wolfowitz Runs Test using Cut Point : 53.889

Variable	Cases <= Cut	Cases > Cut	Runs	Z	p-value (2-tail)
MIL	28.000	28.000	17.000	-3.237	0.001



The test is significant ( $p\text{-value} = 0.001$ ). The military expenditures are not ordered randomly in the file.

The European countries are first in the file, followed by Islamic and New World countries. Looking at the plot, notice that the first 20 cases exceed the median. The remaining cases are for the most part below the median. Iraq, Libya, and Canada stand apart from the other countries in their group. When the line joining the *MIL* values crosses the median line, a new run begins. Thus, the plot illustrates the 17 runs.

## Computation

### Algorithms

Probabilities for the Kolmogorov-Smirnov statistic for  $n < 25$  are computed with an asymptotic negative exponential approximation.

Lilliefors probabilities are computed by a nonlinear approximation to Lilliefors's values. Dallal and Wilkinson (1986) recomputed the Lilliefors's table using up to a million replications for estimating critical values. They found a number of Lilliefors's values to be incorrect. Consequently, the SYSTAT approximation uses the corrected values. The approximation discussed in Dallal and Wilkinson and used in SYSTAT differs from the tabled values by less than 0.01 and by less than 0.001 for  $p < 0.05$ .

For the  $p$ -value associated with the Anderson-Darling test statistic we use formulae from Marsaglia and Marsaglia (2004).

## References

- Anderson, T. W. and Darling, D. A. (1952). Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes, *Annals of Mathematical Statistics*, 23, 193-212.
- Anderson, T. W. and Darling, D. A. (1954). A test of goodness of fit. *Journal of the American Statistical Association*, 49, 765-769.
- Conover, W. J. (1999). *Practical nonparametric statistics*, 3rd ed. New York: John Wiley & Sons.
- Dallal, G.E. and Wilkinson, L. (1986). An analytic approximation to the distribution of Lilliefors' test for normality. *The American Statistician*, 40, 294-296.
- Hollander, M. and Wolfe, D. A. (1999). *Nonparametric statistical methods*, 2nd ed. New York: John Wiley & Sons.
- Lehmann, E. L. and D'Abrera, H. J. M (2006). *Nonparametrics: Statistical methods based on ranks*. New York: Springer-Verlag.

- Marascuilo, L. A. and McSweeney, M. (1977). *Nonparametric and distribution-free methods for the social sciences*. Belmont, Calif.: Wadsworth Publishing.
- Marsaglia, G. and Marsaglia, J. C. W. (2004). Evaluating the Anderson-Darling Distribution. *Journal of Statistical Software*, Vol. 9 -2.
- Mosteller, F. and Rourke, R. E. K. (1973). *Sturdy statistics*. Reading, Mass.: Addison-Wesley.
- Kutner, M. H., Nachtsheim, C. J., Neter, J., and Li, W. (2004) *Applied linear regression models*. 5th ed. New York: McGraw-Hill / Irwin.
- Siegel, S. and Castellan, N.J. (1988). *Nonparametric statistics for the behavioral sciences*, 2nd ed. New York: McGraw-Hill.
- \* Stephens, M. A. (1982). Anderson-Darling test of goodness of fit. *Encyclopedia of Statistical Sciences: Volume 1* (Edited by Kotz, S. and Johnson, N.L). New York: John Wiley & Sons, 81-85.

(\* indicates additional references)



# *Partial Least Squares Regression*

*Moumita Mitra and Suresh Konapalli*

The Partial Least Squares (PLS) technique is one way to construct regression equations; in fact, it can be looked upon as an extension of the multiple linear regression technique. PLS has recently gained importance in many areas of application such as chemometry and economics, especially in situations where the number of variables is large relative to the number of cases, or when there is likely to be multicollinearity among the predictor variables. The PLS method extracts some latent factors from the response and predictor variables separately, and then fits a regression of the response factors on the predictor factors.

SYSTAT offers two of the most popular algorithms for PLS: the Nonlinear Iterative Partial Least Squares (NIPALS) algorithm and the Straight-forward Implementation of Partial Least Squares (SIMPLS) algorithm. The standard errors of the estimated regression coefficients (rather, mean squared errors of these biased estimators) are calculated by the Jackknife procedure. The user is offered two cross-validation procedures, viz., Leave-one-out and Random exclusion, to validate the fitted regression model. SYSTAT provides score plot(s) as Quick Graphs. Further, the coefficient matrix, residuals, predicted values and latent scores can be saved to a SYSTAT file for further analysis.

As cross-validation techniques are available, resampling techniques are not offered in SYSTAT under the PLS regression feature.

## *Statistical Background*

Wold (1966) introduced the PLS technique in the field of econometrics. The use of PLS in chemical applications was pioneered by Wold, Martens and Wold (1983). The

PLS technique is more robust than classical multiple linear regression (univariate or multivariate) and principal component regression. It is robust in the sense that the estimates of model parameters do not change very much when new calibration samples are taken from the population (Geladi and Kowalski, 1986).

When the number of predictors is large, multicollinearity among them is expected. In that case, an ordinary multiple regression technique is unsuitable. Moreover, for a successful application of the multiple regression technique, the number of cases (observations) needs to be much more than the number of variables or the number of parameters to be estimated. PLS is a method intended to alleviate these difficulties in ordinary multiple regression.

## Model Building

The chief purpose of PLS regression is to build a linear model,

$$Y = XB + E^*$$

where  $Y$  is an  $n \times m$  response matrix ( $n$  cases,  $m$  variables),  $X$  is an  $n \times p$  predictor or design matrix ( $n$  cases,  $p$  predictors),  $B$  is a  $p \times m$  matrix of regression coefficients, and  $E^*$  is an  $n \times m$  matrix of noise or error terms.

Usually, before fitting the model, we transform all the observations to a mean-centered or a scaled form in respect of the corresponding variables.

The main approach of PLS is to form components that capture most of the information in the  $X$  variables that is useful to predict  $Y$  variables, while reducing the dimensionality of the regression problem by using fewer components than the number of  $X$  variables (Garthwaite, 1994). In PLS, if the number of extracted factors is greater than or equal to the rank of the  $X$  matrix, then PLS reduces to Multiple Linear Regression.

PLS builds a decomposition of the  $X$  variables as:

$$X = TP' + E = \sum t_h p_h' + E$$

where  $E$  is  $n \times p$ ,  $T$  is  $n \times c$  and  $P$  is  $p \times c$  matrix, with  $c$  being the number of  $X$ -factors. This relation is often called the *outer* relation for  $X$ . A similar outer relation is formed for  $Y$  (see Geladi and Kowalski, 1986):

$$Y = UQ' + F = \sum u_h q_h' + F$$

where  $\mathbf{F}$  is  $n \times m$ ,  $\mathbf{U}$  is  $n \times c$  and  $\mathbf{Q}$  is  $m \times c$  matrix, and  $c$  is the number of  $\mathbf{Y}$ -factors.

In both the cases the summation is taken over  $k=1, 2, \dots, c$ . The matrices  $\mathbf{E}$  and  $\mathbf{F}$  are called the residual matrices. One could take  $c=p$  to make  $\mathbf{E} = \mathbf{F} = 0$ . In general, our intention is to minimize  $\|\mathbf{E}\|$  and  $\|\mathbf{F}\|$ .

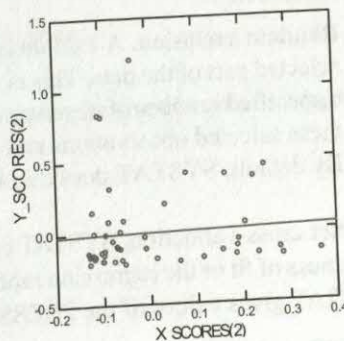
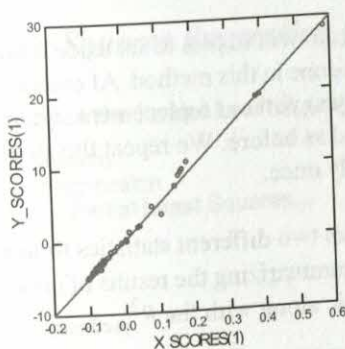
Then a relation between  $\mathbf{X}$  and  $\mathbf{Y}$  (called the inner relation) is developed using the multivariate regression of  $\mathbf{U}$  on  $\mathbf{T}$ .

### Choice of Number of Factors to Extract

A number of latent factors are to be extracted separately from the predictors and the responses. A choice is to be made on the number of such factors. In general, this choice is subjective. It can be any integer between 1 and the rank of the matrix  $\mathbf{X}'\mathbf{X}$ . However, it would be useful to have  $c$  as somewhat smaller than this rank. If  $c$  is too small, then there is a certain loss of information, and if  $c$  is too large, besides the complexity of computation, we may run into the same problems that we seek remedy from PLS. So, there has to be a trade-off between these two aspects to decide on an optimal value of  $c$ .

We can see how well the extracted factors represent the relationship between predictors and responses by plotting  $\mathbf{Y}$ -scores vs.  $\mathbf{X}$ -scores. Let us look at an example of such factors by means of scatter plots of  $\mathbf{X}$ ,  $\mathbf{Y}$  factors.

### Score Plots



The graph on the left is the plot of the first  $\mathbf{X}$ -factor scores vs. the first  $\mathbf{Y}$ -factor scores. There is a high level of correlation between them. Thus the first factor pair derived from the data explains the relation between the predictors and responses very well. On the other hand, the plot of the second  $\mathbf{X}$ -factor scores vs. the second  $\mathbf{Y}$ -factor scores



shows a fairly small level of correlation. Thus the second factor pair does not explain the relation between the predictors and responses well. Generally, the more factors we extract, the less the latter factors are likely to be useful for prediction. In order to decide how many factors are useful, PRESS and  $R^2_{\text{prediction}}$  statistics are used.

## Cross-Validation

Cross-validation is a model evaluation technique that is better than the method using residuals. In residual analysis, we evaluate the efficiency of the model by using the same data which have been used to fit the model. Thus the analysis is likely to be optimistic. The best way to overcome this problem is to use separate datasets for the estimation of the model and for the validation of the model. But, in practical situations we have only one dataset. The most primitive method is the controlled or uncontrolled division of the sample data into two subsamples (Stone, 1977). One part of the dataset is used to fit a model, while the other is used to check the fitting. The first part is known as the 'training set' and the other one as the 'test set'. This is the basic idea for a whole class of model evaluation methods called cross-validation.

SYSTAT offers two types of cross-validation:

- **Leave-one-out.** Here, one observation is removed at each step from the total of  $n$  observations. The remaining  $(n-1)$  observations are used to fit the model and the removed one is used to validate it. This process is repeated  $n$  times, omitting each observation in turn. Mosteller and Tukey (1968) termed this as "simple cross-validation".
- **Random exclusion.** A cautious statistician would like to set aside a randomly selected part of the data. This is what is done in this method. At each step, we select a specified number of observations (say  $s$ ) without replacement. Then we exclude these selected observations and proceed as before. We repeat this process  $r$  times. By default, SYSTAT does this step only once.

After cross-validation, SYSTAT calculates two different statistics to indicate the goodness of fit of the regression model by summarizing the results of cross-validation. SYSTAT gives values of the PRESS statistic along with the  $R^2_{\text{prediction}}$ .

**PRESS statistic.** This statistic is the sum of squares of residuals. Let  $Y_i$  be the observed value of the response for the  $i^{\text{th}}$  individual, and  $\hat{Y}_i$  be the corresponding predicted value. Then the statistic is given by:



$$PRESS = \sum_i (y_i - \hat{y}_i)^2$$

the sum being taken over all the observations under cross-validation.

These PRESS residuals are useful for many purposes. The PRESS statistic is useful for computing the predictive ability of the fitted model which is found by calculating the  $R^2$  statistic for prediction ( $R^2_{\text{prediction}}$ ). This is very similar to the usual  $R^2$  statistic. This statistic is given by:

$$R^2_{\text{prediction}} = 1 - \frac{PRESS}{SST}$$

where, SST is the total sum of squares.

Note that  $R^2_{\text{prediction}}$  lies in the interval [0, 1]: the larger the value of the statistic, the better is the model.

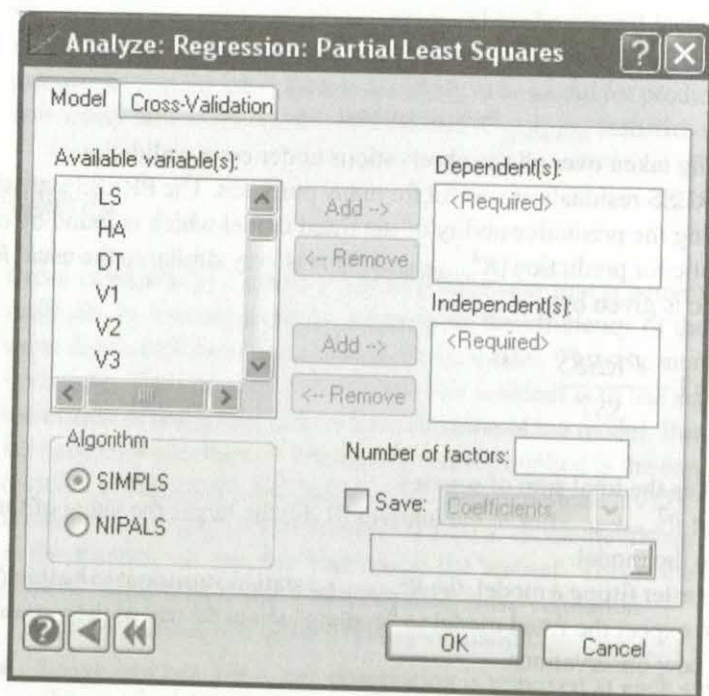
Suppose after fitting a model, the  $R^2_{\text{prediction}}$  statistic turns out to be 0.946. It implies that we can expect the fitted model to "explain" about 94.6% of the variability in predicting new observations.

## ***Partial Least Squares Regression in SYSTAT***

### ***Partial Least Squares Regression Dialog Box***

To open the Partial Least Squares Regression dialog box, from the menus choose:

Analyze  
Regression  
Partial Least Squares...



**Dependent(s).** Select the dependent variable(s) for your study. The dependent variable(s) should be numeric.

**Independent(s).** Select one or more independent variable(s). The independent variable(s) should be numeric.

**Algorithm.** You can choose any of the following algorithms for fitting the data:

- **SIMPLS.** Estimates the model by Straight-forward IMplementation of Partial Least-Squares method. This is the default option.
- **NIPALS.** Estimates the model by Nonlinear Iterative PArTial Least Squares method.

**Number of factors.** Specify the number of latent factors to derive. The number specified should be a positive integer.

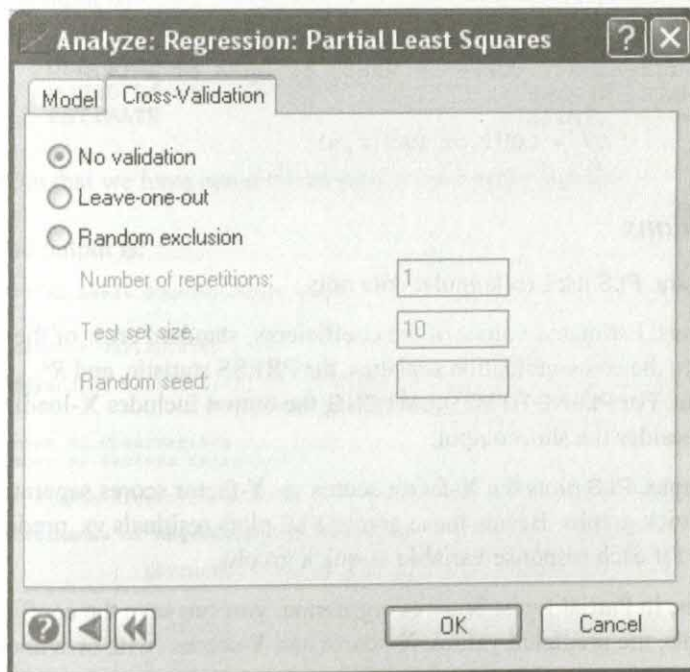
**Save.** Saves the specified results by checking the check-box corresponding to the Save option. The following options are available for saving.

- **Coefficients.** Saves the estimated coefficient matrix. If there is only one response variable, then it is the coefficient vector.

- **Residuals.** Saves the residuals and predicted values.
- **Residuals/data.** Saves all the residuals, predicted values and also the data from the original file.
- **Scores.** Saves **X**-scores and **Y**-scores for all the factors extracted.
- **Scores/data.** Saves **X**-scores, **Y**-scores and also the data from the original file.

### Cross-validation

You can specify different cross-validation options by clicking the Cross-validation tab in the Partial Least Squares regression dialog box.



The following options are available:

**No validation.** No cross-validation is performed.

**Leave-one-out.** The cross-validation is performed by the Leave-one-out technique.

**Random exclusion.** Cross-validation is performed by the random exclusion technique. You can specify the following options:

- **Number of repetitions.** Specify the number of times it should be repeated. By default it is one.
- **Test set size.** Specify the number of observations to be excluded at each step. By default it is half of the total number of observations.
- **Random seed.** Specify any integer from 1 to 4294967295. Otherwise it is based on the system time.

## Using Commands

```
PLS
USE filename
MODEL Y varlist = X varlist / N = n
SAVE filename / COEFF or RESID or DATA or SCORE
ESTIMATE/ NIPALS
SIMPLS
CV = LOU or RAN(r,s)
```

## Usage Considerations

**Types of data.** PLS uses rectangular data only.

**Print options.** Estimated values of the coefficients, standard error of the estimated coefficients, the cross-validation statistics, the PRESS statistic, and  $R^2_{\text{prediction}}$  form the short output. For PLENGTH MEDIUM/LONG, the output includes X-loadings, Y-loadings, besides the short output.

**Quick Graphs.** PLS plots the X-factor scores vs. Y-factor scores separately for each factor as quick graphs. Beside these graphs PLS plots residuals vs. predicted values separately for each response variable as quick graphs.

**Saving files.** In Partial Least Squares regression, you can save the coefficient matrix, the residuals, the predicted values, X-scores and Y-scores (with or without data).

**By groups.** PLS analyzes data by groups.

**Case frequencies.** FREQ is not available in PLS.

**Case weights.** WEIGHT is not available in PLS.



## Examples

### Example 1 Univariate Regression by PLS Technique

We use the SPECTRO data to illustrate the Partial Least Squares Regression. Suppose, we want to predict the amount of Lignin Sulfonate (*LS*) in the Baltic sea with some spectroscopic observations, viz. *V1* to *V27*. We notice that the number of independent variables is larger than the number of observations and so we cannot perform ordinary least-squares regression. Therefore, we perform PLS regression.

The input is:

```
PLS
USE SPECTRO
MODEL LS =V1..V27/N=5
ESTIMATE
```

Note that we have opted for extraction of 5 latent factors.

The output is:

Partial Least Squares Regression

Dependent Variable(s) : LS

Independent Variable(s): V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15  
V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27

Number of Observations : 16  
Number of Factors Extracted : 5

The SIMPLS Algorithm is used to Estimate the Model.

#### Estimates of Regression Coefficients

	ESTIMATE	Standard Error
Constant	0.518261	0.194749
V1	-0.000010	0.000245
V2	-0.000476	0.000187
V3	-0.000111	0.000143
V4	0.000007	0.000075
V5	0.000193	0.000089
V6	0.000267	0.000085
V7	0.000279	0.000088
V8	0.000228	0.000146
V9	0.000087	0.000179
V10	-0.000072	0.000201
V11	-0.000211	0.000162
V12	-0.000358	0.000135
V13	-0.000377	0.000085
V14	-0.000287	0.000052
V15	-0.000269	0.000097

V16	-0.000026	0.000236
V17	0.000058	0.000302
V18	0.000058	0.000282
V19	0.000195	0.000335
V20	0.000290	0.000311
V21	0.000190	0.000118
V22	0.000248	0.000234
V23	0.000334	0.000194
V24	0.000361	0.000189
V25	0.000603	0.000568
V26	0.000770	0.000451
V27	0.000730	0.000600

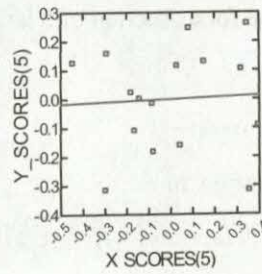
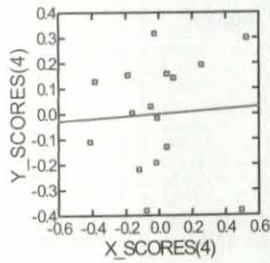
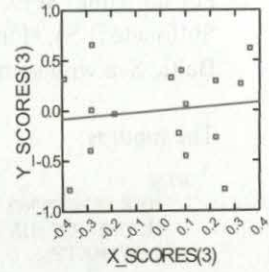
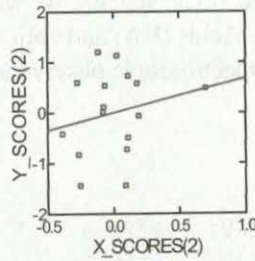
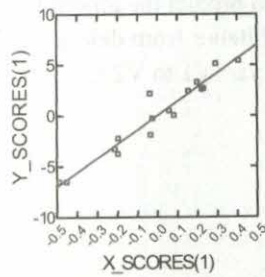
## Analysis of Variance for LS

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	25.073660	5	5.014732	791.198393	0.000000
Error	0.063381	10	0.006338		

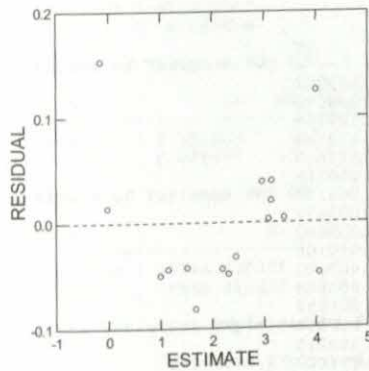
## Percent Variation Explained by Factors for Predictors and Responses

Factors	Variation Explained for Predictor(s)		Variation Explained for Response(s)	
	Percentage	Cum. Percentage	Percentage	Cum. Percentage
1	97.459066	97.459066	93.283983	93.283983
2	2.183130	99.642197	4.563171	97.847154
3	0.145927	99.788124	1.362918	99.210072
4	0.137574	99.925698	0.319522	99.529594
5	0.055504	99.981202	0.218262	99.747856

## Score Plots



Plot of Residuals vs Predicted Values



## Example 2

### Multivariate Regression by PLS Technique

For the same SPECTRO data, suppose we want to predict the amount of Lignin Sulfonate (LS), Humic Acids (HA) and optical whitener from detergent (DT) in the Baltic Sea with some spectroscopic observations, viz. V1 to V27.

The input is:

```
PLS
USE SPECTRO
MODEL LS HA DT=V1..V27/N=5
ESTIMATE
```

Note that we have opted for extraction of 5 latent factors.

The output is:

Partial Least Squares Regression

Dependent Variable(s) : LS HA DT

Independent Variable(s): V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15  
V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27

Number of Observations : 16

Number of Factors Extracted : 5

The SIMPLS Algorithm is used to Estimate the Model.

#### Estimates of Regression Coefficients

	LS	HA	DT
Constant	0.426815	-0.024251	-72.809052
V1	0.000007	-0.000189	0.034156
V2	-0.000446	0.000132	0.027799
V3	-0.000093	-0.000007	0.010768
V4	0.000058	-0.000060	0.003813
V5	0.000173	-0.000113	-0.000312
V6	0.000221	-0.000120	-0.003177
V7	0.000220	-0.000089	-0.005677
V8	0.000195	-0.000033	-0.008600
V9	0.000082	0.000077	-0.010106
V10	-0.000050	0.000177	-0.009080
V11	-0.000161	0.000234	-0.005869
V12	-0.000296	0.000297	-0.001054
V13	-0.000307	0.000284	0.000954
V14	-0.000248	0.000221	0.002215
V15	-0.000281	0.000227	0.004407
V16	-0.000098	0.000107	0.001800
V17	-0.000047	0.000066	0.001790
V18	-0.000021	0.000052	0.001367
V19	0.000052	-0.000008	0.001529
V20	0.000212	-0.000109	-0.001725
V21	0.000256	-0.000153	-0.001145
V22	0.000281	-0.000182	-0.001520
V23	0.000280	-0.000166	-0.002884



## Partial Least Squares Regression

V24	0.000355	-0.000227	-0.002866
V25	0.000815	-0.000526	-0.011565
V26	0.000876	-0.000559	-0.013337
V27	0.000899	-0.000581	-0.012934

## Standard Error of the Estimated Coefficients

	LS	HA	DT
Constant	0.316011	0.161005	57.622733
V1	0.000388	0.000269	0.060957
V2	0.000369	0.000186	0.050844
V3	0.000202	0.000117	0.025321
V4	0.000114	0.000066	0.009864
V5	0.000063	0.000034	0.015814
V6	0.000191	0.000082	0.029646
V7	0.000133	0.000048	0.022328
V8	0.000209	0.000072	0.022533
V9	0.000316	0.000110	0.020193
V10	0.000318	0.000115	0.021109
V11	0.000305	0.000127	0.024917
V12	0.000227	0.000108	0.027643
V13	0.000116	0.000070	0.020389
V14	0.000105	0.000056	0.015533
V15	0.000142	0.000062	0.019092
V16	0.000457	0.000186	0.044954
V17	0.000370	0.000146	0.037095
V18	0.000424	0.000173	0.039827
V19	0.000792	0.000338	0.076063
V20	0.000693	0.000299	0.065644
V21	0.000145	0.000093	0.019493
V22	0.000316	0.000159	0.013095
V23	0.000722	0.000320	0.076940
V24	0.000558	0.000250	0.066951
V25	0.001173	0.000499	0.134145
V26	0.000357	0.000223	0.037164
V27	0.000652	0.000293	0.090291

## Analysis of Variance for LS

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	25.050857	5	5.010171	581.335075	0.000000
Error	0.086184	10	0.008618		

## Analysis of Variance for HA

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	0.552608	5	0.110522	41.460927	0.000002
Error	0.026657	10	0.002666		

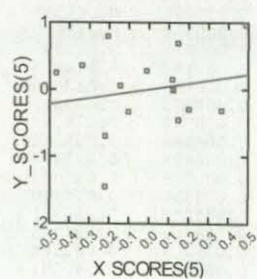
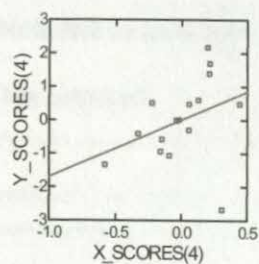
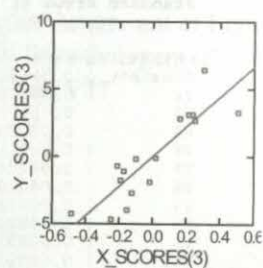
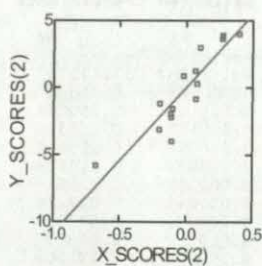
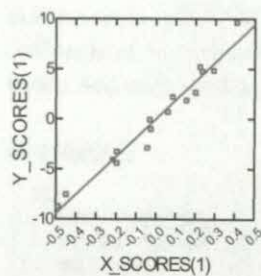
## Analysis of Variance for DT

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	20546.278371	5	4109.255674	21.188084	0.000050
Error	1939.418272	10	193.941827		

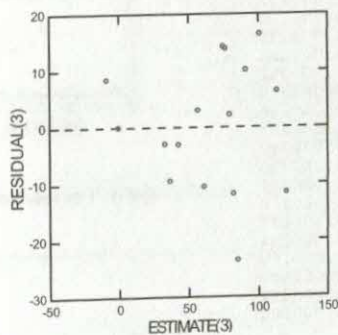
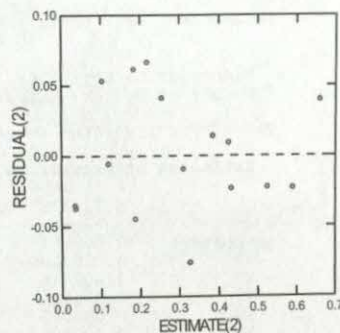
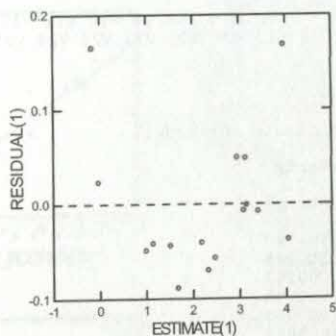
## Percent Variation Explained by Factors for Predictors and Responses

Factors	Variation Explained for Predictor(s)		Variation Explained for Response(s)	
	Percentage	Cum. Percentage	Percentage	Cum. Percentage
1	97.460684	97.460684	41.915455	41.915455
2	2.182960	99.643643	24.243755	66.159210
3	0.177948	99.821591	24.557349	90.716559
4	0.119751	99.941343	3.769548	94.486106
5	0.041593	99.982936	0.990623	95.476730

## Score Plots



### Plot of Residuals vs Predicted Values



### Example 3 Cross-Validation

To assess the fitted model, we can use any one of the cross-validation techniques (say "Leave-one-out") available in PLS.

The input is:

```
PLS
USE SPECTRO
MODEL LS=V1..V27/N=5
ESTIMATE/CV=LOUT
```

The output is:

Partial Least Squares Regression

Dependent Variable(s) : LS

Independent Variable(s): V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15  
V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27

Number of Observations : 16

Number of Factors Extracted : 5

The SIMPLS Algorithm is used to Estimate the Model.

#### Estimates of Regression Coefficients

	ESTIMATE	Standard Error
Constant	0.518261	0.194749
V1	-0.000010	0.000245
V2	-0.000476	0.000187
V3	-0.000111	0.000143
V4	0.000007	0.000075
V5	0.000193	0.000089
V6	0.000267	0.000085
V7	0.000279	0.000088
V8	0.000228	0.000146
V9	0.000087	0.000179
V10	-0.000072	0.000201
V11	-0.000211	0.000162
V12	-0.000358	0.000135
V13	-0.000377	0.000085
V14	-0.000287	0.000052
V15	-0.000269	0.000097
V16	-0.000026	0.000236
V17	0.000058	0.000302
V18	0.000058	0.000282
V19	0.000195	0.000335
V20	0.000290	0.000311
V21	0.000190	0.000118
V22	0.000248	0.000234
V23	0.000334	0.000194
V24	0.000361	0.000189
V25	0.000603	0.000568
V26	0.000770	0.000451
V27	0.000730	0.000600

#### Analysis of Variance for LS

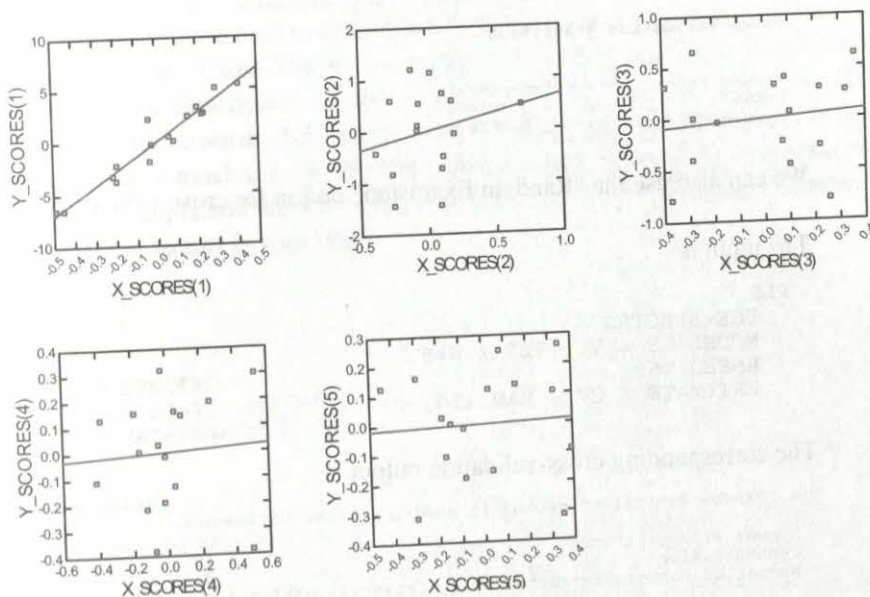
Source	SS	df	Mean Squares	F-ratio	p-value
Regression	25.073660	5	5.014732	791.198393	0.000000
Error	0.063381	10	0.006338		

#### Percent Variation Explained by Factors for Predictors and Responses

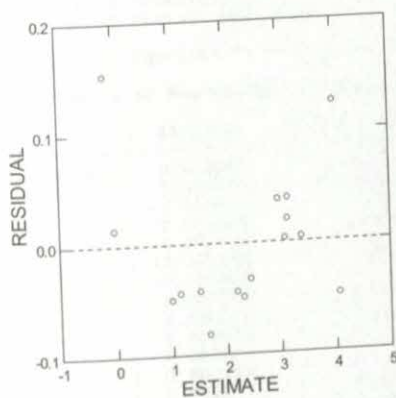
Factors	Variation Explained for Predictor(s)		Variation Explained for Response(s)	
	Percentage	Cum.Percentage	Percentage	Cum.Percentage
1	97.459066	97.459066	93.283983	93.283983
2	2.183130	99.642197	4.563171	97.847154
3	0.145927	99.788124	1.362918	99.210072
4	0.137574	99.925698	0.319522	99.529594
5	0.055504	99.981202	0.218262	99.747856



## Score Plots



Plot of Residuals vs Predicted Values



The "Leave One Out" method is used for Cross-Validation.

Number of Factors Extracted after Cross-Validation : 5

**Cross-Validation Statistics**

	LS
PRESS	0.451494
R-square(Prediction)	0.982039

We can also use the "Random Exclusion" option for cross-validation.

The input is:

```
PLS
USE SPECTRO
MODEL LS = V1..V27 / N=5
RSEED 459
ESTIMATE / CV = RAN (10, 9)
```

The corresponding cross-validation output is:

The "Random Exclusion" method is used for Cross-Validation.

```
Number of Repetitions      : 10
Test Set Size              : 9
Number of Factors Extracted after Cross-Validation : 5
```

**Cross-Validation Statistics**

	LS
Average PRESS	0.581181
R-square(Prediction)	0.976879

### Example 4

#### Optimum Choice of Number of Factors

We now demonstrate how to determine the optimum number of factors. Look at the score plots in the "Cross-Validation" example. We can see that the X-scores and the Y-scores are very closely, and linearly, related for the first and second factors. But from the third factor onwards, they become more or less dispersed. So the last three factors are not really of much use. We can thus repeat the same analysis by extracting only two factors. The explained variance table also indicates that the explained variance due to the last three factors is very small.

The input is:

```
PLS
USE SPECTRO
MODEL LS=V1..V27/N=2
ESTIMATE/CV=LOUT
```

The output is:

Partial Least Squares Regression

Dependent Variable(s) : LS

Independent Variable(s): V1 V2 V3 V4 V5 V6 V7 V8 V9 V10 V11 V12 V13 V14 V15  
V16 V17 V18 V19 V20 V21 V22 V23 V24 V25 V26 V27

Number of Observations : 16  
Number of Factors Extracted : 2

The SIMPLS Algorithm is used to Estimate the Model.

#### Estimates of Regression Coefficients

	ESTIMATE	Standard Error
Constant	-0.003824	0.200336
V1	-0.000140	0.000046
V2	-0.000087	0.000044
V3	-0.000035	0.000034
V4	-0.000019	0.000025
V5	-0.000008	0.000023
V6	-0.000001	0.000022
V7	0.000005	0.000021
V8	0.000014	0.000018
V9	0.000022	0.000016
V10	0.000028	0.000011
V11	0.000031	0.000009
V12	0.000031	0.000007
V13	0.000032	0.000007
V14	0.000040	0.000008
V15	0.000045	0.000012
V16	0.000056	0.000014
V17	0.000069	0.000016
V18	0.000083	0.000022
V19	0.000102	0.000029
V20	0.000129	0.000034

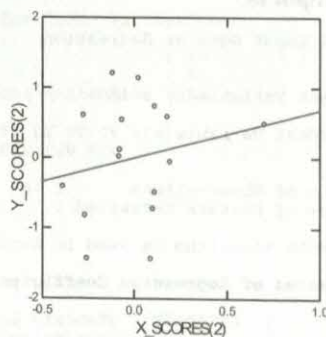
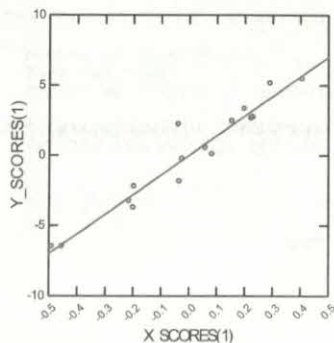
V21	0.000153	0.000042
V22	0.000182	0.000043
V23	0.000207	0.000049
V24	0.000237	0.000064
V25	0.000269	0.000063
V26	0.000301	0.000074
V27	0.000321	0.000079

**Analysis of Variance for LS**

Source	SS	df	Mean Squares	F-ratio	p-value
Regression	24.595879	2	12.297940	295.425923	0.000000
Error	0.541162	13	0.041628		

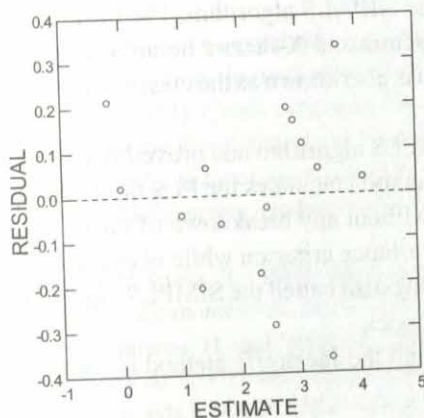
**Percent Variation Explained by Factors for Predictors and Responses**

Factors	Variation Explained for Predictor(s)		Variation Explained for Response(s)	
	Percentage	Cum. Percentage	Percentage	Cum. Percentage
1	97.459066	97.459066	93.283983	93.283983
2	2.183130	99.642197	4.563171	97.847154

**Score Plots**



Plot of Residuals vs Predicted Values



The "Leave One Out" method is used for Cross-Validation.

Number of Factors Extracted after Cross-Validation : 2

#### Cross-Validation Statistics

	LS
PRESS	0.973167
R-square (Prediction)	0.961286

Note that the X and Y scores for the two factors extracted are more or less linearly related. As far as the goodness of the model is concerned we can say that we have not lost much by reducing the number of factors. (Note that the  $R^2$  prediction of the second model is 0.961286 compared to 0.982039 of first model.) We can therefore say that the extraction of two factors is satisfactory.

## Computation

### Algorithms

SYSTAT provides two options for fitting the Partial Least Squares regression model: SIMPLS (Straight-forward Implementation of Partial Least Squares) and NIPALS (Nonlinear Iterative Partial Least Squares).

The NIPALS algorithm was proposed by H. Wold (1966) in the context of estimation of principal components in Multivariate Analysis. Geladi and Kowalski (1986) gave a clear exposition of the NIPALS algorithm. The basic idea of the algorithm is to find orthogonal **Y**-factors and **X**-factors iteratively by deflating the centered data matrix. This algorithm is also known as the classical algorithm of the PLS method.

de Jong (1993) proposed the SIMPLS algorithm and proved that it is better than the original classical algorithm. This method calculates the PLS factors directly as linear combinations of original variables without any breakdown of the dataset. Factors are determined so as to maximize a covariance criterion while obeying the orthogonality and normalization restrictions. de Jong also called the SIMPLS algorithm Statistically Inspired Method of Partial Least Squares.

The standard error calculation using the Jackknife method is a time-consuming exercise.

## Missing Data

SYSTAT deletes missing values by the list-wise deletion technique, i.e., it ignores those cases which have at least one missing value (whether in response or in predictors).

## References

- \* Burnham, A.J., MacGregor, J.F., and Viveris, R. (1999). Latent variable regression tools. *Chemometrics and Intelligent Laboratory Systems*, 48, 167-180.
- de Jong, S. (1993). SIMPLS: an alternative approach to partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 18, 251-263.
- \* Denham, M. C. (1997). Prediction intervals in partial least squares. *Journal of Chemometrics*, 11, 39-52.
- Garthwaite, P.H. (1994). An interpretation of partial least squares. *Journal of the American Statistical Association*, 89, 122-127.
- Geladi, P., and Kowalski B.R. (1986). Partial least-squares regression: A tutorial. *Analytica Chimica Acta*, 185, 1-17.
- \* Montgomery D.C., Peck E.A., and Vining G.G. (2006). *Introduction to linear regression analysis*. 4th ed. Hoboken, N.J.: Wiley-Interscience.
- Mosteller, F. and Tukey, J.W. (1968). Data analysis including statistics. *Handbook of*

- Social Psychology* (G. Lindzey and E. Aronson, eds). Reading, Mass. Addison-Wesley.
- \* Srivastava M.S. (2002). *Methods of multivariate statistics*. New York: John Wiley & Sons.
- Stone, M. (1977). An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. *Journal of the Royal Statistical Society. B*, 39, 44-47.
- \* Stone, M. (1974). Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society. B*, 36, 111-147.
- Wold, H. (1966). Estimation of principal components and related models by iterative least squares. *Multivariate Analysis* (P.R. Krishnaiah, ed.), 391-420. New York: Academic Press.
- Wold, H. and Strotz, R. (1960). Recursive versus nonrecursive systems: an attempt at synthesis. *Econometrica*, 28, 417-427.
- Wold, S., Martens, H. and Wold, H. (1983). The Multivariate Calibration Problem in Chemistry solved by the PLS Method. Proc. Conf. Matrix Pencils, (A. Ruhe and B. Kågström, eds.), March 1982. *Lecture Notes in Mathematics*, Springer Verlag, Heidelberg, 286-293.

(\* indicating additional reference)

## Statistical Background





# *Partially Ordered Scalogram Analysis with Coordinates*

*Leland Wilkinson, Samuel Shye, Reuben Amar, and Louis Guttman*

The POSAC module calculates a partial order scalogram analysis on a set of multicategory items. It consolidates duplicate data profiles, computes profile similarity coefficients, and iteratively computes a configuration of points in a two-dimensional space according to the partial order model. POSAC produces Quick Graphs of the configuration, labeled by either profile values or an ID variable. Shye (1985) is the authoritative reference on POSAC. See also Borg's review (1987) for more information. The best approach to set up a study for POSAC analysis is to use facet theory (see Canter, 1985).

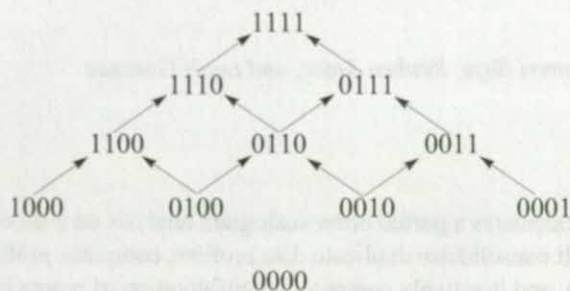
Resampling procedures are available in this feature.

## *Statistical Background*

The following figure shows a pattern of bits in two dimensions, an instance of a partially ordered set (POSET). There are several interesting things about this pattern.

- The vertical dimension of the pattern runs from four 1's on the top to no 1's on the bottom.
- The horizontal dimension runs from 1's on the left to 1's in the center to 1's on the right.
- Except for the bottom row, each bit pattern is the result of an OR operation of the two bit patterns below itself, as denoted by the arrows in the figure. For example, (1111) = (1110) or (0111).

- There are  $2^4 = 16$  possible patterns for four bits. Only 11 patterns meet the above requirements in two dimensions. The remaining patterns are: (1011), (1101), (1010), (0101), and (1001).
- This structure is a *lattice*. We can move things around and still represent the POSET geometrically as long as none of the arrows cross or head down instead of up.



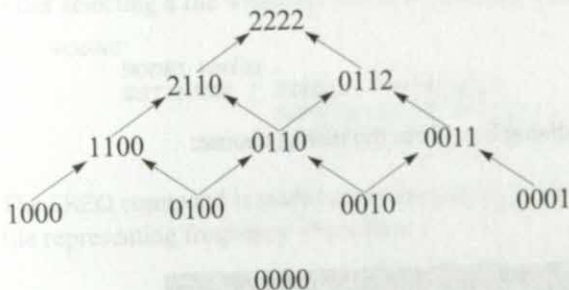
Suppose we had real binary data involving the presence or absence of attributes and wanted to determine whether our data fit a POSET structure. We would have to do the following:

- Order the attributes from left to right so that the horizontal dimension would show 1's moving from left to right in the plotted profile, as in the figure above.
- Sort the profiles of attributes from top to bottom.
- Sort the profiles from left to right.
- Locate any profiles not fitting the pattern and make sure the overall solution was not influenced by them.

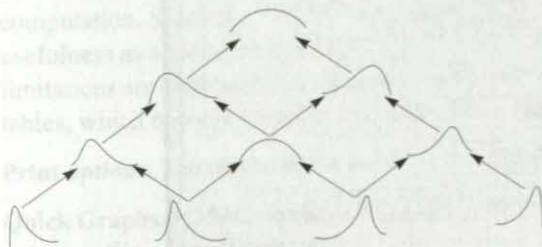
The fourth requirement is somewhat elusive and depends on the first. That is, if we had patterns (1010) and (0101), exchanging the second and third bits would yield (1100) and (0011), which would give us two extreme profiles in the third row rather than two ill-fitting profiles. If we exchange bits for one profile, we must exchange them for all, however. Thus, the global solution depends on the order of the bits as well as their positioning.

POSAC stands for partially ordered scalogram analysis with coordinates. The algorithm underlying POSAC computes the ordering and the lattice for cases-by-attributes data. Developed originally by Louis Guttman and Samuel Shye, POSAC fits, not only binary but also multivalued, data into a two-dimensional space according to the constraints we have discussed.

The following figure (a multivalued POSET) shows a partial ordering on some multivalued profiles. Again, we see that the marginal values increase on the vertical dimension (from 0 to 1 to 2 to 4 to 8) and the horizontal dimension distinguishes left and right skew.



The following figure shows this distributional positioning more generally. For ordered profiles with many values on each attribute, we expect the central profiles in the POSAC to be symmetrically distributed, profiles to the left to be right-skewed, and profiles to the right to be left-skewed.



### Coordinates

There are two standard coordinate systems for displaying profiles. The first uses joint and lateral dimensions to display the profiles as in the figures above. Profiles that have similar sum scores fall at approximately the same latitude in this coordinate system. Comparable profiles differing in their sum scores (for example, 112211 and 223322) fall above and below each other at the same longitude.

The second coordinate display, the one printed in the SYSTAT plots, is a 45-degree rotation of this set. These base coordinates have the joint dimension running from

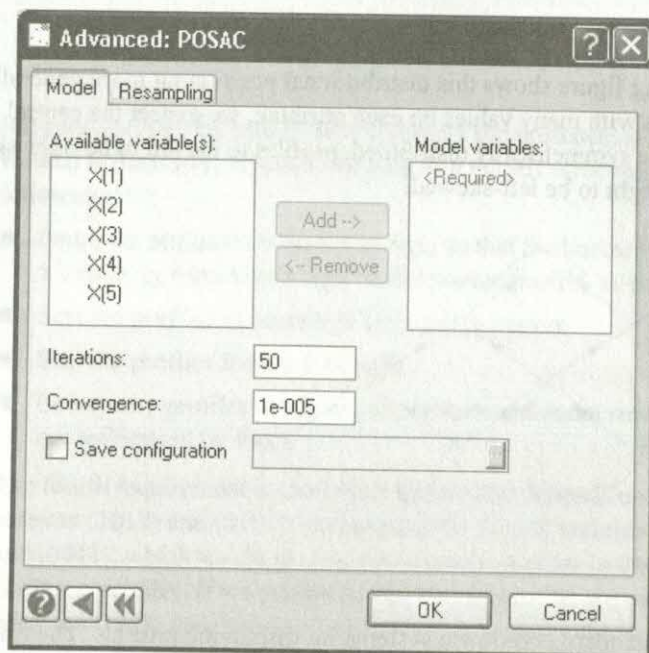
southwest to northeast and the lateral dimension running from northwest to southeast. The diamond pattern is transformed into a square.

## POSAC in SYSTAT

### POSAC Dialog Box

To open the POSAC dialog box, from the menus choose:

Advanced  
POSAC...



**Model variables.** Specify the items to be scaled. Select at least three items.

**Iterations.** Enter the maximum number of iterations that you wish to allow the program to perform in order to estimate the parameters.

**Convergence.** Enter the convergence criterion. This is the largest relative change in any coordinate before iterations terminate.



**Save configuration.** You can save the configuration into a SYSTAT file.

## Using Commands

After selecting a file with **USE filename**, continue with:

```
POSAC
  MODEL varlist
  ESTIMATE / ITER=n, CONVERGE=d
             SAMPLE =BOOT (m,n) or
             SIMPLE (m,n) or JACK
```

The **FREQ** command is useful when data are aggregated and there is a variable in the file representing frequency of profiles.

## Usage Considerations

**Types of data.** POSAC only uses rectangular data. It is most suited for data with up to nine categories per item. If your data have more than nine categories, the profile labels will not be informative, since each item is displayed with a single digit in the profile labels. If your data have many more categories in an item, the program may refuse the computation. Similarly, POSAC can handle many items, but its interpretability and usefulness as an analytical tool declines after 10 or 20 items. These practical limitations are comparable to those for loglinear modeling and analysis of contingency tables, which become complex and problematic for multiway tables.

**Print options.** The output is the same for all **PLENGTH** options.

**Quick Graphs.** POSAC produces a Quick Graph of the coordinates labeled either with value profiles or an ID variable.

**Saving files.** POSAC saves the configuration into a file.

**BY groups.** POSAC analyzes data by groups. Your file need not be sorted on the **BY** variable(s).

**Case frequencies.** **FREQ <variable>** increases the number of cases by the **FREQ** variable.

**Case weights.** **WEIGHT** is not available in POSAC.

## Examples

The following examples illustrate the features of the POSAC module. The first example involves binary profiles that fit the POSAC model perfectly. The second example shows an analysis for real binary data. The third example shows how POSAC works for multicategory data.

### Example 1

#### Scalogram Analysis—A Perfect Fit

The file *BIT5* contains five-item binary profiles fitting a two-dimensional structure perfectly.

The input is:

```
USE BIT5
POSAC
MODEL X(1) .. X(5)
ESTIMATE
```

The output is:

#### Partially Ordered Scalogram

##### Reordered Item Weak Monotonicity Coefficients

	X(5)	X(4)	X(3)	X(2)	X(1)
X(5)	1.000				
X(4)	0.750	1.000			
X(3)	0.111	0.667	1.000		
X(2)	-0.286	0.000	0.667	1.000	
X(1)	-0.391	-0.286	0.111	0.750	1.000

#### Iteration History

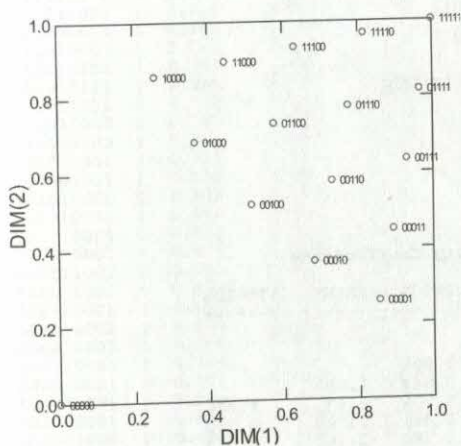
Iteration	Loss
1	0.017
2	0.007
3	0.002
4	0.000
5	0.000
6	0.000

```
Final Loss Value : 0.000
Proportion of Profile Pairs Correctly Represented : 1.000
Score-distance Weighted Coefficient : 1.000
```

LABEL\$	DIM(1)	DIM(2)	JOINT	LATERAL	FIT
11111	1.000	1.000	1.000	0.500	0.000
01111	0.966	0.816	0.891	0.575	0.000
11110	0.816	0.966	0.891	0.425	0.000
01110	0.775	0.775	0.775	0.500	0.000

00111	0.931	0.632	0.782	0.649	0.000
11100	0.632	0.931	0.782	0.351	0.000
01100	0.577	0.730	0.654	0.424	0.000
00011	0.894	0.447	0.671	0.724	0.000
11000	0.447	0.894	0.671	0.276	0.000
00110	0.730	0.577	0.654	0.576	0.000
00100	0.516	0.516	0.516	0.500	0.000
10000	0.258	0.856	0.557	0.201	0.000
00010	0.683	0.365	0.524	0.659	0.000
00001	0.856	0.258	0.557	0.799	0.000
01000	0.365	0.683	0.524	0.341	0.000
00000	0.000	0.000	0.000	0.500	0.000

POSAC Profile Plot



POSAC first computes Guttman monotonicity coefficients and orders the corresponding matrix using an SSA (multidimensional scaling) algorithm. These monotonicity coefficients, which Shye (1985) discusses in detail, are similar to the MU2 coefficients in the SYSTAT CORR module.

The next section of the output shows the iteration history and computed coordinates. SYSTAT's POSAC module calculates the square roots of the coordinates before display and plotting. This is done in order to make the lateral direction linear rather than curvilinear. Notice that for the perfect data in this example, the profiles are confined to the upper right triangle of the plot, as in the theoretical examples in Shye (1985). If you are comparing output with the earlier Jerusalem program, remember to include this transformation. Notice that the profiles are ordered in both the joint and lateral directions.

## Example 2

### Binary Profiles

The following data are reports of fear symptoms by selected United States soldiers after being withdrawn from World War II combat. The data were originally reported by Suchman in Stouffer et al. (1950). Notice that we use FREQ to represent duplicate profiles.

The input is:

```
USE COMBAT
FREQ COUNT
POSAC
MODEL POUNDING..URINE
ESTIMATE
```

The output is:

#### Partially Ordered Scalogram

##### Reordered Item Weak Monotonicity Coefficients

	STIFF	VOMIT	NAUSEOUS	FAINT	SINKING
STIFF	1.000				
VOMIT	0.682	1.000			
NAUSEOUS	0.728	0.815	1.000		
FAINT	0.716	0.665	0.844	1.000	
SINKING	0.583	0.381	0.706	0.644	1.000
SHAKING	0.829	0.495	0.661	0.729	0.705
BOWELS	0.751	0.780	0.780	0.761	0.513
URINE	0.782	0.589	1.000	0.846	1.000
POUNDING	0.290	0.443	0.615	0.569	0.449

##### Reordered Item Weak Monotonicity Coefficients (contd...)

	SHAKING	BOWELS	URINE	POUNDING
SHAKING	1.000			
BOWELS	0.617	1.000		
URINE	0.763	0.960	1.000	
POUNDING	0.709	1.000	1.000	1.000

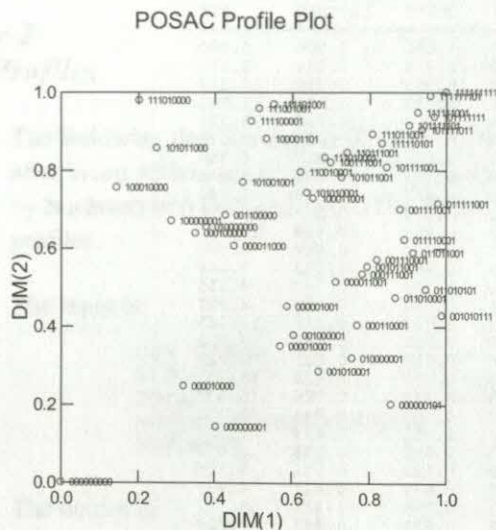
#### Iteration History

Iteration	Loss
1	4.612
2	2.260
3	1.194
4	0.878
5	0.898
Final Loss Value	: 0.878
Proportion of Profile Pairs Correctly Represented	: 0.810
Score-distance Weighted Coefficient	: 0.977



## Partially Ordered Scalogram Analysis with Coordinates

LABELS	DIM(1)	DIM(2)	JOINT	LATERAL	FIT
111111111	1.000	1.000	1.000	0.500	0.000
111111101	0.958	0.990	0.974	0.484	2.577
101111111	0.969	0.937	0.953	0.516	10.242
111111001	0.926	0.948	0.937	0.489	11.973
111110101	0.833	0.869	0.851	0.482	13.251
101111101	0.904	0.915	0.909	0.494	7.571
101111011	0.937	0.904	0.920	0.517	9.357
111101001	0.553	0.969	0.761	0.292	8.880
011111001	0.979	0.714	0.847	0.633	6.641
101111001	0.845	0.808	0.827	0.519	10.411
111011001	0.808	0.892	0.850	0.458	11.101
110111001	0.742	0.845	0.794	0.449	8.689
011110001	0.892	0.623	0.757	0.635	7.238
001111001	0.881	0.700	0.790	0.590	4.255
100111001	0.700	0.821	0.760	0.440	6.911
111001001	0.515	0.958	0.737	0.278	12.063
011011001	0.915	0.589	0.752	0.663	9.030
111100001	0.495	0.926	0.710	0.285	10.225
111010001	0.685	0.833	0.759	0.426	13.307
011010101	0.948	0.495	0.721	0.726	5.937
001010111	0.990	0.429	0.709	0.781	1.793
101011001	0.728	0.782	0.755	0.473	8.332
101011000	0.247	0.857	0.552	0.195	8.936
111010000	0.202	0.979	0.591	0.111	9.716
001011001	0.795	0.553	0.674	0.621	4.639
100001101	0.535	0.881	0.708	0.327	18.117
101010001	0.639	0.742	0.691	0.448	6.088
011010001	0.869	0.474	0.671	0.698	10.334
001110001	0.821	0.571	0.696	0.625	7.902
110010001	0.623	0.795	0.709	0.414	9.413
000111001	0.782	0.535	0.658	0.624	6.454
101001001	0.474	0.769	0.622	0.352	6.892
100011001	0.655	0.728	0.692	0.463	7.752
001010001	0.670	0.286	0.478	0.692	7.128
000011001	0.714	0.515	0.615	0.600	8.843
000110001	0.769	0.404	0.587	0.683	7.337
000010001	0.571	0.350	0.461	0.611	1.155
100000001	0.286	0.670	0.478	0.308	5.579
001000001	0.606	0.378	0.492	0.614	7.827
000011000	0.452	0.606	0.529	0.423	9.295
001100000	0.429	0.685	0.557	0.372	10.533
100010000	0.143	0.756	0.449	0.193	10.084
000001001	0.589	0.452	0.520	0.569	8.718
000000101	0.857	0.202	0.530	0.828	15.543
010000001	0.756	0.319	0.538	0.718	18.115
000000001	0.404	0.143	0.273	0.631	6.259
000100000	0.350	0.639	0.494	0.356	10.401
010000000	0.378	0.655	0.516	0.362	13.698
000010000	0.319	0.247	0.283	0.536	11.087
000000000	0.000	0.000	0.000	0.500	0.000



The output shows an initial ordering of the symptoms that, according to the SSA, runs from stiffness to loss of urine and bowel control and a pounding heart. The lateral dimension follows this general ordering. Notice that the joint dimension runs from absence of symptoms to presence of all symptoms.

### Example 3

#### Multiple Categories

This example uses crime data to construct a 2D solution of crime patterns. We first recode the data into four categories for each item by using the CUT function. The cuts are made at each standard deviation and the mean. Then, POSAC computes the coordinates for these four category profiles.

The input is:

```
USE CRIME
STANDARDIZE MURDER..AUTOTHFT
LET (MURDER..AUTOTHFT)=CUT(@,-1,0,1,4)
POSAC
MODEL MURDER..AUTOTHFT
ESTIMATE
```

# Partially Ordered Scalogram Analysis with Coordinates

The output is:

## Partially Ordered Scalogram

### Reordered Item Weak Monotonicity Coefficients

	LARCENY	AUTOTHFT	BURGLARY	ROBBERY	RAPE
LARCENY	1.000				
AUTOTHFT	0.821	1.000			
BURGLARY	0.930	0.950	1.000		
ROBBERY	0.806	0.900	0.868	1.000	
RAPE	0.786	0.731	0.850	0.922	1.000
ASSAULT	0.516	0.667	0.742	0.879	0.921
MURDER	0.280	0.483	0.579	0.650	0.823

### Reordered Item Weak Monotonicity Coefficients (contd...)

	ASSAULT	MURDER
ASSAULT	1.000	
MURDER	0.965	1.000

### Iteration History

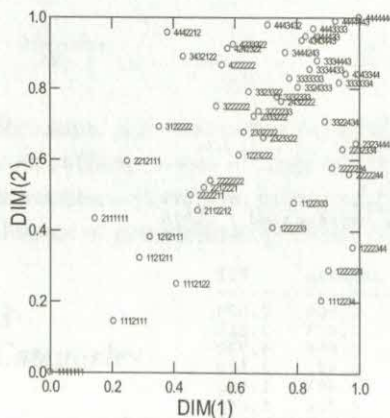
Iteration	Loss
1	0.451
2	0.333
3	0.131
4	0.102
5	0.085
6	0.091

Final Loss Value : 0.085  
 Proportion of Profile Pairs Correctly Represented : 0.816  
 Score-distance Weighted Coefficient : 0.994

LABELS	DIM(1)	DIM(2)	JOINT	LATERAL	FIT
4444444	1.000	1.000	1.000	0.500	0.000
4444443	0.924	0.990	0.957	0.467	2.015
4343344	0.957	0.842	0.900	0.558	4.770
4344433	0.829	0.946	0.888	0.441	2.576
4343443	0.816	0.935	0.876	0.441	1.995
4443432	0.707	0.979	0.843	0.364	1.045
4443333	0.854	0.968	0.911	0.443	2.559
3444243	0.764	0.901	0.833	0.431	3.171
3334443	0.866	0.878	0.872	0.494	1.569
3334433	0.842	0.854	0.848	0.494	1.148
3333334	0.935	0.816	0.876	0.559	2.027
2323444	0.990	0.645	0.818	0.672	0.437
3333333	0.777	0.829	0.803	0.474	0.563
3324333	0.804	0.804	0.804	0.500	3.832
3322434	0.890	0.707	0.798	0.591	4.147
3332333	0.736	0.777	0.757	0.479	2.577
4442212	0.382	0.957	0.670	0.212	2.154
4233322	0.595	0.924	0.760	0.335	3.045
2232334	0.946	0.629	0.788	0.659	0.692
4242322	0.577	0.913	0.745	0.332	2.624
2222244	0.968	0.559	0.764	0.705	2.340
1222344	0.979	0.354	0.666	0.813	2.170
3323322	0.645	0.791	0.718	0.427	1.750
3432122	0.433	0.890	0.661	0.272	4.266
2323322	0.692	0.661	0.677	0.515	2.677
2333222	0.661	0.722	0.692	0.470	2.352
2222234	0.913	0.577	0.745	0.668	1.941
3222233	0.677	0.736	0.706	0.471	2.052
2432222	0.750	0.764	0.757	0.493	6.825
2332222	0.629	0.677	0.653	0.476	2.881
4222222	0.559	0.866	0.713	0.346	0.920

1122333	0.791	0.479	0.635	0.656	4.239
3222222	0.540	0.750	0.645	0.395	1.711
1222233	0.722	0.408	0.565	0.657	2.231
1222224	0.901	0.289	0.595	0.806	1.819
1223222	0.612	0.612	0.612	0.500	6.108
1112234	0.878	0.204	0.541	0.837	1.259
2222222	0.520	0.540	0.530	0.490	1.193
3122222	0.354	0.692	0.523	0.331	5.871
2222211	0.456	0.500	0.478	0.478	2.515
2212221	0.500	0.520	0.510	0.490	2.936
2112212	0.479	0.456	0.468	0.511	3.532
2212111	0.250	0.595	0.423	0.327	2.841
1112122	0.408	0.250	0.329	0.579	2.135
1212111	0.323	0.382	0.352	0.470	2.938
1121211	0.289	0.323	0.306	0.483	3.621
2111111	0.144	0.433	0.289	0.356	3.497
1112111	0.204	0.144	0.174	0.530	0.309
1111111	0.000	0.000	0.000	0.500	0.000

POSAC Profile Plot



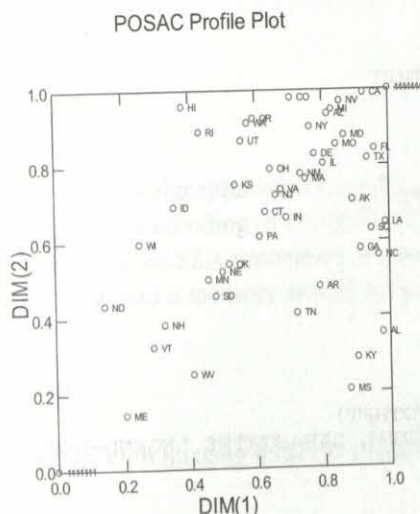
The configuration plot is labeled with the profile values. We can see that the larger values generally fall in the upper extreme of the joint (diagonal) dimension. The lateral dimension runs basically according to the ordering of the initial SSA, from property crimes at the left end of each profile to person crimes at the right end. POSAC thus has organized the states in two dimensions by frequency (low versus high) and by type of crime (person versus property).

If we add

```
IDVAR STATES
```

before the ESTIMATE command, we can label the points with the state names. The result is shown in the following POSAC profile plot:





### POSAC and MDS

To see how the POSAC compares to a multidimensional scaling (MDS), we ran an MDS on the transposed crime data. The following input program illustrates several important points about SYSTAT and data analyses in this context. Our goal is to run an MDS on the distances (differences) between states on crime incidence for the seven crimes. First, we standardize the variables so that all of the crimes have a comparable influence on the differences between states. This prevents a high-frequency crime, like auto theft, from unduly influencing the crime differences. Next, we add a *LABELS* variable to the file because *TRANPOSE* renames the variables with its values if a variable with this name is found in the source file. We save the transposed file into *TCRIME* and then use *CORR* to compute Euclidean distances between the states. MDS is then used to analyze the matrix of pairwise distances of the states ranging from Maine to Hawaii (the two-letter state names are from the U.S. Post Office designations).

We save the MDS configuration instead of looking at the plot immediately because we want to do one more thing. We are going to make the symbol sizes proportional to the standardized level of the crimes (by summing them into a *TOTAL* crime variable). States with the highest value on this variable rank highest, in general, on all crimes. By merging *SCRIME* (produced by the original standardization) and *CONF* (produced by MDS), we retain the labels, the crime values and the configuration coordinates.

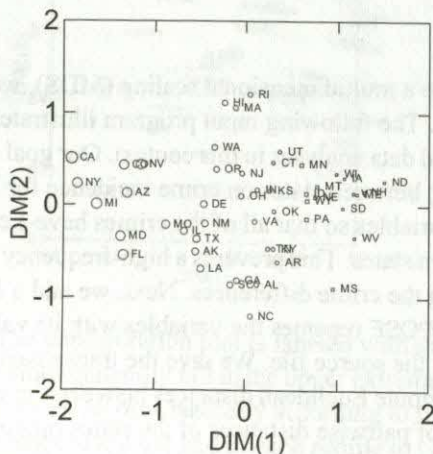
The input is:

```

USE CRIME
STANDARDIZE MURDER..AUTOTHFT
DSAVE SCRIME
CORR
USE SCRIME
LET LABEL$=STATE$
TRANSPose MURDER..AUTOTHFT
SAVE TCRIME
EUCLID ME..HI
MDS
USE TCRIME
MODEL ME..HI
SAVE CONF / CONFIG
ESTIMATE
MERGE CONF SCRIME
LET TOTAL=SUM(MURDER..AUTOTHFT)
PLOT DIM(2)*DIM(1)/SIZE=TOTAL, LAB=STATE$, LEGEND=NONE

```

The output is:



Notice that the first dimension comprises a frequency of crime factor since the size of the symbols is generally larger on the left. This dimension is not much different from the joint dimension in the POSAC configuration. The second dimension, however, is less interpretable than the POSAC lateral dimension. It is not clearly person versus property.

## Computation

### Algorithms

POSAC uses algorithms developed by Louis Guttman and Samuel Shye. The SYSTAT program is a recoding of the Hebrew University version using different minimization algorithms, an SSA procedure to reorder the profiles according to a suggestion of Guttman, and a memory model which allows large problems.

### Missing Data

Profiles with missing data are excluded from the calculations.

## References

- Borg, I. (1987). Review of S. Shye, Multiple scaling. *Psychometrika*, 52, 304–307.
- \* Borg, I. and Shye, S. (1995). *Facet theory: Form and content*. Thousand Oaks, Calif.: Sage Publications.
- Canter, D. [Ed]. (1985). *Facet theory approaches to social research*. New York: Springer Verlag.
- \* Shye, S. [Ed]. (1978). *Theory construction and data analysis in the behavioral sciences*. San Francisco, Calif.: Jossey-Bass.
- Shye, S. (1985). *Multiple scaling: The theory and application of partial order scalogram analysis*. Amsterdam: North-Holland.
- Stouffer, S. A., Guttman, L., Suchman, E. A., Lazarsfeld, P. F., Staf, S. A., and Clausen, J. A. (1950). *Measurement and prediction*. Princeton, N.J.: Princeton University Press.


(\* indicates additional references)





# *Path Analysis (RAMONA)*

Michael W. Browne



RAMONA implements the McArdle and McDonald Reticular Action Model (RAM) for path analysis with manifest and latent variables. Input to the program is coded directly from a path diagram without reference to any matrices.

RAMONA stands for *RAM Or Near Approximation*. The deviation from RAM is minor—no distinction is made between residual variables and other latent variables. As in RAM, only two parameter matrices are involved in the model. One represents single-headed arrows in the path diagram (path coefficients) and the other, double-headed arrows (covariance relationships).

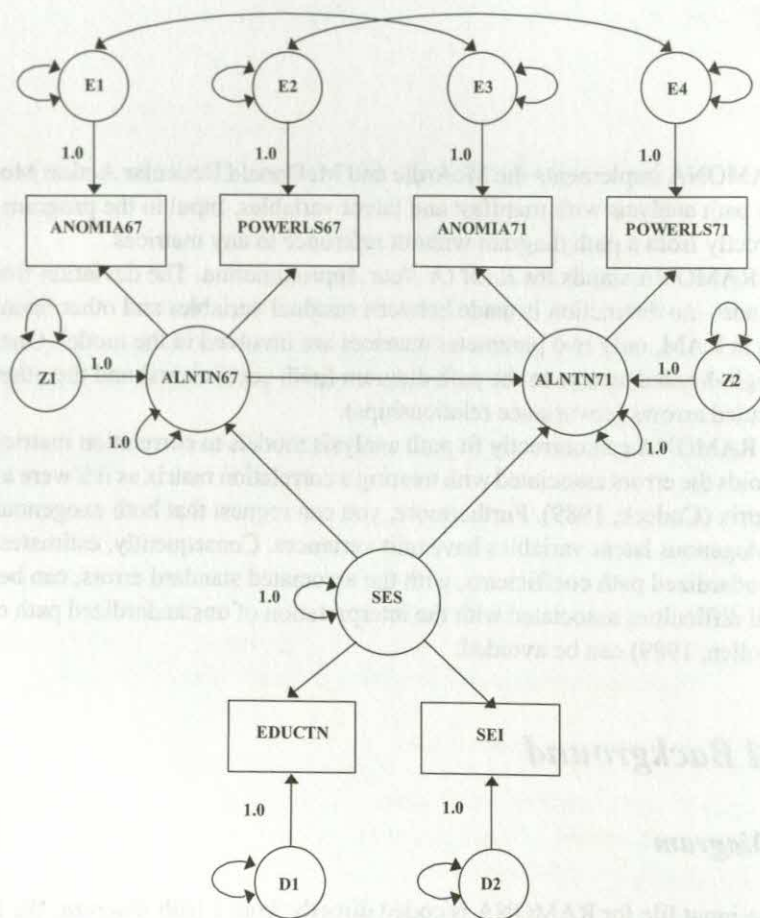
RAMONA can correctly fit path analysis models to correlation matrices, and it avoids the errors associated with treating a correlation matrix as if it were a covariance matrix (Cudeck, 1989). Furthermore, you can request that both exogenous and endogenous latent variables have unit variances. Consequently, estimates of standardized path coefficients, with the associated standard errors, can be obtained, and difficulties associated with the interpretation of unstandardized path coefficients (Bollen, 1989) can be avoided.

## *Statistical Background*

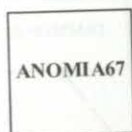
### *The Path Diagram*

The input file for RAMONA is coded directly from a path diagram. We first briefly review the main characteristics of path diagrams. More information can be found in texts dealing with structural equation modeling (Bollen, 1989; Everitt, 1984; and McDonald, 1985).

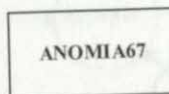
Look at the following path diagram. This is a model, adapted from Jöreskog (1977), for a study of the stability of attitudes over time conducted by Wheaton, Muthén, Alwin, and Summers (1977). Attitude scales measuring anomia (*ANOMIA*) and powerlessness (*POWERLS*) were regarded as indicators of the latent variable alienation (*ALNTN*) and administered to 932 persons in 1967 and 1971. A socioeconomic index (*SEI*) and years of school completed (*EDUCTN*) were regarded as indicators of the latent variable socioeconomic status (*SES*).



In the path diagram, a manifest (observed) variable is represented by a square or rectangular box:



or



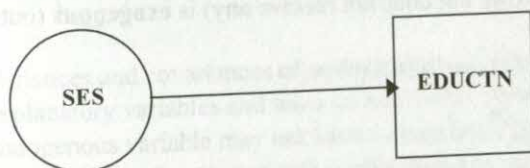
while a circle or ellipse signifies a latent (unobservable) variable:



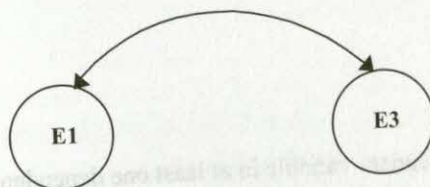
or



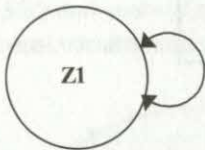
A **dependence path** is represented by a single-headed arrow emitted by the *explanatory* variable and received by the *dependent* variable:



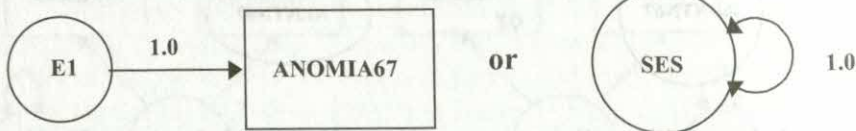
while a **covariance path** is represented by a double-headed arrow:



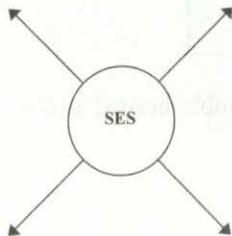
In many diagrams, **variance paths** are omitted. Because variances form an essential part of a model and must be specified for RAMONA, we represent them here explicitly by curved double-headed arrows (McArdle, 1988) with both heads touching the same circle or square:



If a path coefficient, variance, or covariance is fixed (at a nonzero value), we attach the value to the single- or double-headed arrow:

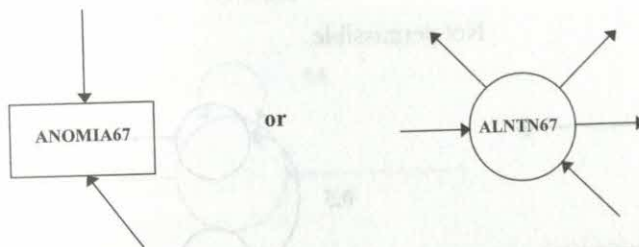


A variable that acts as an explanatory variable in all of its dependence relationships (emits single-headed arrows but does not receive any) is **exogenous** (outside the system):



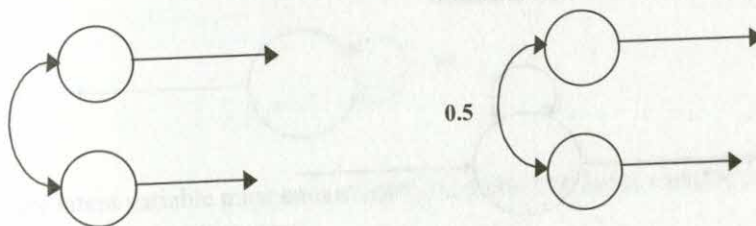
A variable that acts as a dependent variable in at least one dependence relationship (receives at least one single-headed arrow) is **endogenous** (inside the system), whether or not it ever acts as an explanatory variable (emits any arrows):





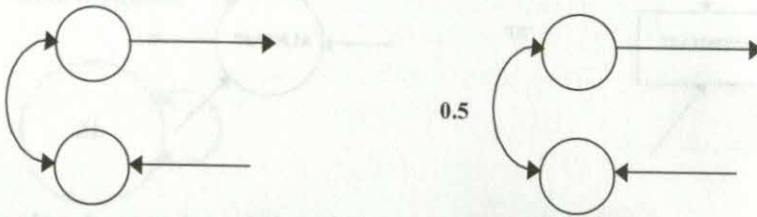
A parameter in RAMONA is associated with each dependence path and covariance path between two exogenous variables. Covariance paths are permitted only between exogenous variables. For example, the following covariance paths are permissible:

Permissible



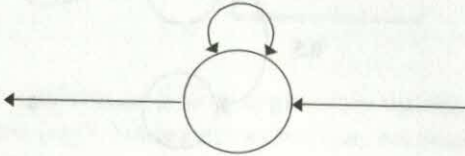
Variances and covariances of endogenous variables are implied by the corresponding explanatory variables and have no associated parameters in the model. Thus, an endogenous variable may not have a covariance path with any other variable. The covariance is a function of path coefficients and variances or covariances of exogenous variables and is not represented by a parameter in the model. The following covariance paths, for example, are not permissible:

Not permissible



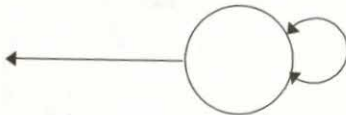
Also, an endogenous variable does not have a free parameter representing its variance. Its variance is a *function* of the path coefficients and variances of its explanatory variables. Therefore, it may not have an associated double-headed arrow with no fixed value:

Not Permissible



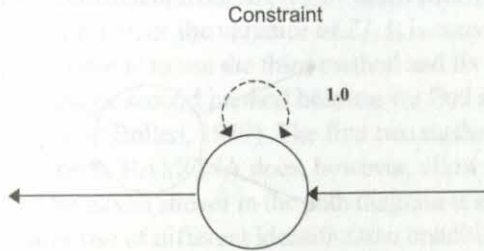
Exogenous variables alone may have free parameters representing their variances:

Permissible



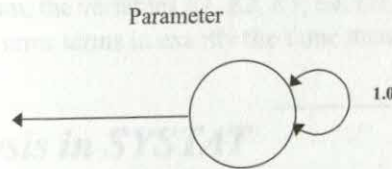
We do, however, allow *fixed* variances for both endogenous and exogenous variables. These two types of fixed variances are interpreted differently in the program:

- A fixed variance for an endogenous variable is treated as a nonlinear equality constraint on the parameters in the model:

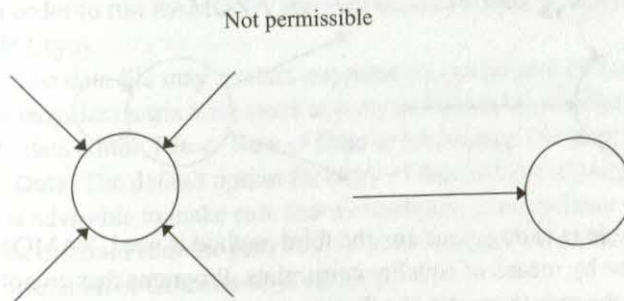


The fixed implied variance is represented by a dotted two-headed arrow instead of a solid two-headed arrow because it is a nonlinear constraint on several other parameters in the model and does not have a single fixed parameter associated with it.

- A fixed variance for an exogenous variable is treated as a model parameter with a fixed value:



Every latent variable must emit at least one arrow. No latent variable can receive arrows without emitting any:

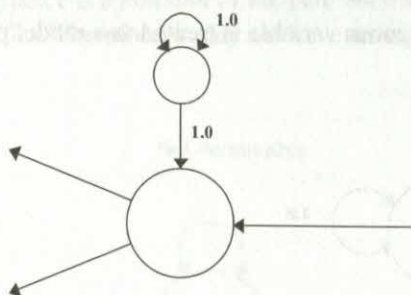


The scale of every latent variable (exogenous or endogenous) should be fixed to avoid indeterminate parameter values. Some ways for accomplishing this are:

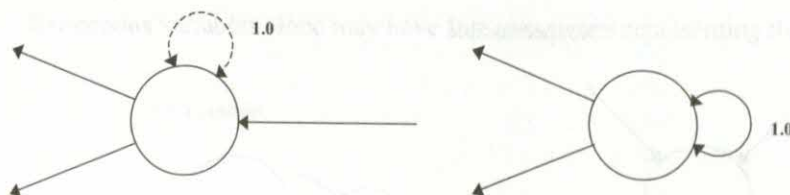
- To fix one of the path coefficients, associated with an emitted arrow, to a nonzero value (usually 1.0):



- To fix both the variance and the path coefficient of an associated error term, if the latent variable is endogenous:



- To fix the variance of the latent variable:



If a latent variable is endogenous and the third method is used, RAMONA fixes the implied variance by means of equality constraints. Programs that do not have this facility require the user to employ the first or second method to determine the scales of endogenous latent variables.

Consider *ALNTN67* in the path diagram. This latent variable is endogenous (it receives arrows from *SES* and *ZI*). It also emits arrows to *ANOMIA67* and *POWRLS67*. Consequently, it is necessary to fix either the variance of *ALNTN67*, the



path coefficient from *ALNTN67* to *ANOMIA67*, the path coefficient from *ALNTN67* to *POWRLS67*, or the variance of *Z1*. It is conventional to use 1.0 as the fixed value. Our preference is to use the third method and fix the variance of *ALNTN67* rather than use the first or second method because we find standardized path coefficients easier to interpret (Bollen, 1989). The first two methods result in latent variables with non-unit variances. RAMONA does, however, allow the use of these methods.

The model shown in the path diagram is equivalent to Jöreskog's (1977) model but makes use of different identification conditions. We apply nonlinear equality constraints to fix the variances of the endogenous variables *ALNTN67* and *ALNTN71*, but treat the path coefficients from *ALNTN67* to *ANOMIA67* and from *ALNTN71* to *ANOMIA71* as free parameters. Jöreskog fixed the path coefficients from *ALNTN67* to *ANOMIA67* and from *ALNTN71* to *ANOMIA71* and did not apply any nonlinear equality constraints.

An error term is an exogenous latent variable that emits only one single-headed arrow and shares double-headed arrows only with other error terms. In the path diagram, the variables *E1*, *E2*, *E3*, *E4*, *D1*, *D2*, *Z1*, and *Z2* are error terms. RAMONA treats error terms in exactly the same manner as other latent variables.

## Path Analysis in SYSTAT

### Instructions for using RAMONA

In order to run RAMONA you will need two files: a data file (.syd) and a command file (.syc).

The data file may contain a symmetric covariance or correlation matrix or a rectangular matrix with cases as rows and variables as columns. It may be entered with the data editor, File -> New -> Data or an existing file may be employed, File -> Open -> Data. The default option for entry of data is for a rectangular matrix. Consequently it is advisable to make sure that a correlation or covariance matrix is not specified as a data matrix. From the path File -> Save As, click on Options and ensure that Correlation or Covariance is selected.

The command file gives a full specification of the analysis to be carried out. To create a new command file click File -> New -> Command and enter the statements. To save the command file click File -> Save As and provide a file name.

An example of a path diagram follows. It represents the Wheaton-Muthen-Alwin-Summers model shown in the path diagram in the section headed The Path Diagram

```

RAMONA
USE EX1
TITLE 'Wheaton, Muthen, Alwin and Summers (1977)
Example'
MANIFEST ANOMIA67 POWRLS67 ANOMIA71 POWRLS71 EDUCTN
SEI
LATENT ALNTN67 ALNTN71 SES E1 E2 E3 E4 D1 D2 Z1 Z2
MODEL ANOMIA67 <- ALNTN67 E1(0, 1.0) ,
POWRLS67 <- ALNTN67 E2(0, 1.0) ,
ANOMIA71 <- ALNTN71 E3(0, 1.0) ,
POWRLS71 <- ALNTN71 E4(0, 1.0) ,
EDUCTN <- SES D1(0, 1.0) ,
SEI <- SES D2(0, 1.0) ,
ALNTN67 <- SES Z1(0, 1.0) ,
ALNTN71 <- ALNTN67 SES Z2(0, 1.0) ,
SES <-> SES(0, 1.0) ,
E1 <-> E1 E3 ,
E2 <-> E2 E4 ,
E3 <-> E3 ,
E4 <-> E4 ,
D1 <-> D1 ,
D2 <-> D2 ,
Z1 <-> Z1 ,
Z2 <-> Z2 ,
ALNTN71 <-> ALNTN71(0, 1.0) ,
ALNTN67 <-> ALNTN67(0, 1.0)
PLENGTH MEDIUM
ESTIMATE / DISP=CORR METHOD=MWL NCASES=932,
START=ROUGH CONVG=0.0001 ITER=500 CONFI=.90

```

Note that the input is not case sensitive so that lower and/or upper case symbols may be used as desired. RAMONA replaces all lower case names by their upper case equivalents before output.

A brief introduction to the statements in the command file follows.

The first statement "RAMONA" instructs SYSTAT which program to use.

The next statement "USE EX1" specifies the data set to be used.

The next statement "TITLE ..." provides a title for the job (optional).

The next statement "MANIFEST ..." lists the names of the manifest variables, represented in squares in the path diagram. (Optional but recommended)

The next statement "LATENT ..." lists the names of the latent variables, represented in circles in the path diagram. (Optional but recommended).

The following three statements "MODEL ..." for specifying the model from the path diagram, "PLENGTH ..." for specifying the amount of output required and "ESTIMATE ..." will be described in detail in the subsections that follow:

## *The MODEL statement*

### **Dependence Relationships**

Each single headed arrow (dependence path) in the path diagram must be indicated by a statement with the symbol <-. To code a dependence path, enter the descriptive name of the dependent variable followed by the symbol <-. Then name the explanatory variable, followed by two symbols in parentheses separated by a comma, for example:

```
ANOMIA67 <- ALNTN67(1, 0.6)
```

If the first symbol in parentheses is a positive integer, 1 in this example, it refers to the parameter group number. The parameter associated with it is constrained to be equal to every other parameter with the same parameter group number. Thus the regression paths associated with

```
ANOMIA67 <- ALNTN67(1, 0.6)
ANOMIA67 <- ALNTN67(1, 0.6)
```

will have the same value of the associated parameter, or regression weight even though the paths are not the same. If the first symbol in parentheses is an asterisk\* then the parameter is regarded as a free parameter that is not constrained to be equal to any other parameter and is not constrained to have any specified value. If the first symbol in parentheses is a 0 then the parameter is regarded as fixed with the value assigned by the second symbol, which should be a real number, for example 3.6.

The second symbol in parentheses specifies a starting value for the parameter associated with the corresponding path. If the first symbol in parentheses is an asterisk or positive integer the second symbol specifies a starting value that will change during the course of iteration. If the first symbol is a 0 then the second symbol specifies a path equal to the value of the second symbol. If the second symbol is an asterisk then the starting value is chosen by the program. Thus:

- (0, 1.0) specifies a path fixed to 1.0



- (\*, 1.0) specifies a starting value of 1.0 that will change during iteration
- (5, 1.0) represents a starting value for all parameters in group 5.

Contradictory specifications (5, 1.0) and (5, 2.0) should be avoided. Also specifications (0,\*) assigning an unspecified value to a parameter should be avoided. The (\*, \*) specification is a default and may be omitted. Thus "ANOMIA67 <- ALNTN67(\*, \*)" and "ANOMIA67 <- ALNTN67" mean the same thing.

An inspection of the path diagram in the "Statistical Background" section shows that the endogenous manifest variable *POWRLS67* receives single-headed arrows from the latent variable *ALNTN67* and the measurement error *Z1*. These dependence relationships can be coded as:

```
POWRLS67 <- ALNTN67(*, *) ,
POWRLS 67 <- E2(0,1.0)
```

In the first path, the parameter is free and not constrained to equality with any other parameter. The parameter number is replaced by an asterisk\*. No starting value is specified either; this too is replaced by an asterisk\*. The parameter in the second path is fixed at 1.0 so that the parameter number is 0 and the parameter value is 1.0.

It is not necessary to have a different statement for each path. Several paths with the same dependent (receiving) variable can be combined into one statement. Since the same endogenous variable, *POWRLS67*, is involved in two dependence relationships, the two paths can be coded in a single statement as:

```
POWRLS67 <- ALNTN67 E2(0,1.0)
```

Suppose that it is known from a previous run that the path coefficient of *ALNTN67* to *ALNTN71* is approximately 0.6. In this case, you can specify the following:

```
ALNTN71 <- ALNTN67(*,0.6) SES(7,*) Z2(0,1.0)
```

When specifying dependence relationships, bear in mind that:

- Dependence relationships can be specified in any order.
- A statement can specify several dependence paths involving the same dependent variable.
- Specified path numbers need not be sequential; for example, 5, 3, 9 can be used. Sequential path numbers will be reassigned by the program.



### Covariance Relationships

A variance or a covariance relationship is indicated by the symbol  $\leftrightarrow$ , which relates directly to the double-headed arrow in the path diagram. To specify a covariance path, enter the name of one of the variables in the path, followed by the symbol  $\leftrightarrow$ . Then enter the name of the other variable, and include the path number and the starting value within parentheses. Unlike the dependence relationship, it does not matter which variable is given first. For example,

```
E2  $\leftrightarrow$  E2(10,*)
```

Other conventions, however, are similar to those for dependence relationships. You can replace the number and/or the starting value of a free parameter with the symbol \*. In this case, they are provided by the program. In the case of a fixed parameter, however, you must specify 0 as the number of the parameter and provide the fixed value of the parameter. An inspection of the path diagram shows that double-headed arrows are used from the measurement error  $E1$  to itself to specify a variance and to  $E3$  to specify a covariance. These relationships are specified in the statement:

```
E1  $\leftrightarrow$  E1(*,*) E3(*,*)
```

or

```
E1  $\leftrightarrow$  E1 E3
```

The same covariance should not be specified twice. Thus the statement  $E1 \leftrightarrow E3$  should not be duplicated with  $E3 \leftrightarrow E1$ . Covariance paths can be constrained to be equal in the same manner as dependence paths. Suppose you want to specify that the variances of the measurement errors  $E1$ ,  $E2$ , and  $E3$  must be equal:

```
E1  $\leftrightarrow$  E1(10,*) E3,  
E2  $\leftrightarrow$  E2(10,*) ,  
E3  $\leftrightarrow$  E3(10,*)
```

You can again provide starting values for free parameters:

```
E3  $\leftrightarrow$  E3(*,0.32)
```

Variances of both exogenous and endogenous variables can be required to have fixed values. Thus, both

```
SES  $\leftrightarrow$  SES(0,1.0)
```

and

```
ALTNTN67 <-> ALTNTN67(0,1.0)
```

are acceptable. They are, however, treated differently within the program. The exogenous latent variable, *SES*, has a parameter associated with its variance and it is set equal to 1.0. There is no parameter representing the variance of the endogenous latent variable, *ALNTN67*. This variance is a function of the path coefficient, *ALNTN67* <- *SES*, the variance of *SES*, and the variance of *ZI*. It is constrained to have a value of 1.0 by *RAMONA*.

When specifying covariance relationships, bear in mind that:

- Covariance paths can be specified in any order.
- Several covariance paths per statement can be specified. For example, the variance of an exogenous variable as well as its covariances with other exogenous variables can be specified in the same statement.
- The same covariance should not be specified twice. Thus the statement *E1 <-> E3* should not be duplicated with *E3 <-> E1*.
- Dependence paths and covariance paths must be specified in separate substatements. The dependence path subparagraph must precede the covariance path subparagraph.
- If every manifest endogenous variable has a corresponding measurement error with an unconstrained variance, the coding of these variances can be omitted. When all error path coefficients are fixed and no error variance paths are input for the measurement errors, the program will automatically provide the error variance paths.
- If there are exogenous manifest variables and if all of their variances and covariances are present in the system and are unrestricted, the coding of these variance and covariance paths can be omitted. When no variance and covariance paths for exogenous manifest variables are entered, the program will automatically provide them.

The *MODEL* statement will typically consist of a number of lines. It is important to remember to have a comma at the end of every line except the last line.

## RAMONA Options

### The PLENGTH statement

Three lengths of output are available and are specified with the PLENGTH statement.

- PLENGTH SHORT. The sample covariance (correlation) matrix, path coefficient estimates, 90% confidence intervals, standard errors and t statistics, and variance/covariance or correlation estimates.
  - PLENGTH MEDIUM. The panels listed for SHORT, plus details of the iterative procedure, the reproduced covariance or correlation matrix, the matrix of residuals, and information about equality constraints on variances (if applicable).
  - PLENGTH LONG. The panels listed for MEDIUM, plus the asymptotic correlation matrix of the estimators.
- PLENGTH MEDIUM is recommended for general use.

### The ESTIMATE statement

This statement is of the form

ESTIMATE /

It is followed by some or all of the following statements in arbitrary order:

DISP = This specifies the type of dispersion matrix to be analysed.

If DISP = COV (Default) an analysis appropriate for a covariance matrix is carried out. If the input matrix is a correlation matrix (has unit diagonal elements), an analysis appropriate for a covariance matrix is performed, but RAMONA prints a warning in the output.

If DISP = CORR an analysis appropriate for a correlation matrix is carried out. If a covariance matrix has been input from the data file it is converted to a correlation matrix before any analysis is carried out.

Note that if this option is not correctly specified some results provided by the program will be incorrect. See Cudeck (1989).

METHOD = This specifies the method of estimation used.

If METHOD = MWL (Default) Maximum Wishart likelihood estimates are obtained.



If METHOD = GLS Generalized least squares estimates appropriate for a Wishart distribution are obtained.

If METHOD = OLS Ordinary least squares estimates are obtained. No measures of fit and no standard errors of estimators are provided.

If METHOD = ADFG Asymptotically distribution-free estimates are provided. These use a biased but Gramian (non-negative definite) estimate of the asymptotic covariance matrix of sample covariances.

If METHOD = ADFU Asymptotically distribution-free estimates are provided. These use an unbiased estimate of the asymptotic covariance matrix of sample covariances.

NCASES = The number of cases used to compute the covariance or correlation matrix must be provided (e.g. NCASES=932) unless a rectangular data matrix has been provided in the data file. The number of cases should exceed the number of  $p$  manifest variables if you use the maximum Wishart likelihood method or the generalized least squares method. If you use the ADF Gramian method or the ADF unbiased method the number of cases must exceed  $0.5 p(p + 1)$ .

START = This designates how starting values are to be scaled for all estimation methods.

If START = ROUGH, starting values are assumed to be inaccurate. They are rescaled so as to yield an implied dispersion matrix with diagonal elements equal to those of the input dispersion matrix. RAMONA applies ordinary least-squares initially. After partial convergence, RAMONA switches to the method you specify. If you are not sure about the starting values you specify, or if you are using the \* option because the starting values are poor, you are advised to use this option.

If START = CLOSE, RAMONA uses the estimation procedure specified under the Method from the beginning of the iterative procedure. This option should always be used with OLS.

CONVG = A convergence criterion is provided. If the default of CONVG = 0.0001 is used, results will be accurate to about three decimal places.

ITER = The maximum number of iterations is provided. The default is ITER = 100

CONFI = The coverage probability for all confidence intervals is provided. The default is CONFI = 0.9.



## Running RAMONA

RAMONA may be run either by reading the command file to the Command Window (File -> Open -> Command) and executing (File -> Submit -> Window) or by executing directly (File -> Submit -> File). The output may then be printed or saved to a file from the output pane (see output in the SYSTAT index).

## Usage Considerations

**Types of data.** RAMONA uses a correlation or covariance matrix either read from a file or computed from a rectangular file. When specifying ADFG or ADFU, a cases-by-variables input file must be used.

**Print options.** Three lengths of output are available. You can specify using PLENGTH:

- **PLENGTH SHORT.** The sample covariance (correlation) matrix, path coefficient estimates, 90% confidence intervals, standard errors and *t* statistics, and variance/covariance or correlation estimates.
- **PLENGTH MEDIUM.** The panels listed for SHORT, plus details of the iterative procedure, the reproduced covariance or correlation matrix, the matrix of residuals, and information about equality constraints on variances (if applicable).
- **PLENGTH LONG.** The panels listed for MEDIUM, plus the asymptotic correlation matrix of the estimators.

**Quick Graphs.** RAMONA produces no Quick Graphs.

**Saving files.** You cannot save specific RAMONA results to a file.

**BY groups.** For a rectangular file, RAMONA produces separate results for each BY variable.

**Case frequencies.** RAMONA uses a FREQUENCY variable, if present, to duplicate cases.

**Case weights.** RAMONA ignores WEIGHT variables.

## Examples

### Example 1

#### Path Analysis Basics

The covariance matrix of six manifest variables is shown below. These covariances and variances were computed from a sample of 932 respondents and are stored in the EX1 data file.

	ANOMIA67	POWRLS67	ANOMIA71	POWRLS71	EDUCTN	SEI
ANOMIA67	11.834					
POWRLS67	6.947	9.364				
ANOMIA71	6.819	5.091	12.532			
POWRLS71	4.783	5.028	7.495	9.986		
EDUCTN	-3.839	-3.889	-3.841	-3.625	9.610	
SEI	-21.899	-18.831	-21.748	-18.755	35.522	450.288

In this example, we specify the model illustrated in "Statistical Background" on p. 397. The command file is listed in the section "Instructions for using RAMONA" on page 405. The role of the manifest and latent variables is clear from the MODEL statement below. Manifest variables are in the SYSTAT file (latent variables are not).

We use the default maximum Wishart likelihood method (METHOD = MWL) to analyze the correlation matrix. Our analysis differs from Jöreskog's analysis in that the model is treated as a correlation structure rather than a covariance structure. The display correlation option of ESTIMATE (TYPE = CORR) identifies that the input is a correlation matrix, and NCASES = 932 denotes the sample size used to compute it.

The output is:

There are 6 Manifest Variables in the Model. They are  
ANOMIA67 POWRLS67 ANOMIA71 POWRLS71 EDUCTN SEI

There are 11 Latent Variables in the Model. They are  
ALNTN67 E1 E2 ALNTN71 E3 E4 SES D1 D2 Z1 Z2

#### RAMONA Options in Effect are

Display		Corr
Method		MWL
Start		Rough
Convergence Limit		0.0001
Maximum Iterations		100
N of Cases		932
Restart		No
% Confidence Level		90

Number of Manifest Variables : 6  
 Total Number of Variables in the System : 23

Reading Covariance Matrix...

#### Details of Iterations

Iteration	Method	Discr. Funct.	Max.R.Cos.	Max.Const.	NRP	NBD
0	OLS	2.990		0.000		
1(0)	OLS	9363.180	0.999	87.020	0	0
1(1)	OLS	67.826	0.974	9.357	0	0
1(2)	OLS	1.861	0.657	1.221	0	0
2(0)	OLS	0.864	0.645	0.787	0	0
3(0)	OLS	0.020	0.512	0.131	0	0
4(0)	OLS	0.001	0.302	0.004	0	0
5(0)	OLS	0.001	0.001	0.000	0	0
5(0)	MWL	0.005	0.034	0.000	0	0
6(0)	MWL	0.005	0.009	0.000	0	0
7(0)	MWL	0.005	0.001	0.000	0	0
8(0)	MWL	0.005	0.000	0.000	0	0
9(0)	MWL	0.005	0.000	0.000	0	0
10(0)	MWL	0.005	0.000	0.000	0	0

Iterative procedure complete.

Convergence Limit for Residual Cosines : 1.000E-04 on 2 Consecutive Iterations  
 Convergence Limit for Variance Constraint Violations: 5.000E-07  
 Value of the Maximum Variance Constraint Violations : 1.293E-11

#### Sample Correlation Matrix

	ANOMIA67	POWRLS67	ANOMIA71	POWRLS71	EDUCTN	SEI
ANOMIA67	1.000					
POWRLS67	0.660	1.000				
ANOMIA71	0.560	0.470	1.000			
POWRLS71	0.440	0.520	0.670	1.000		
EDUCTN	-0.360	-0.410	-0.350	-0.370	1.000	
SEI	-0.300	-0.290	-0.290	-0.280	0.540	1.000

Number of Cases : 932

#### Reproduced Correlation Matrix

	ANOMIA67	POWRLS67	ANOMIA71	POWRLS71	EDUCTN	SEI
ANOMIA67	1.000					
POWRLS67	0.660	1.000				
ANOMIA71	0.560	0.469	1.000			
POWRLS71	0.441	0.520	0.670	1.000		
EDUCTN	-0.367	-0.404	-0.357	-0.369	1.000	
SEI	-0.280	-0.308	-0.272	-0.281	0.540	1.000

#### Residual Matrix (correlations)

	ANOMIA67	POWRLS67	ANOMIA71	POWRLS71	EDUCTN	SEI
ANOMIA67	0.000					
POWRLS67	0.000	0.000				
ANOMIA71	0.000	0.001	0.000			
POWRLS71	-0.001	0.000	0.000	0.000		
EDUCTN	0.007	-0.006	0.007	-0.001	0.000	
SEI	-0.020	0.018	-0.017	0.001	0.000	0.000



Value of the Maximum Absolute Residual : 0.020

### ML Estimates of Free Parameters in Dependence Relationships

Path	Parameter Number	Point Estimate	90.00% Confidence Interval	
			Lower	Upper
ANOMIA67 <- ALNTN67	1	0.774	0.733	0.816
POWRLS67 <- ALNTN67	2	0.852	0.810	0.894
ANOMIA71 <- ALNTN71	3	0.805	0.763	0.848
POWRLS71 <- ALNTN71	4	0.832	0.788	0.876
EDUCTN <- SES	5	0.842	0.789	0.894
SEI <- SES	6	0.642	0.592	0.691
ALNTN67 <- SES	7	-0.563	-0.620	-0.506
ALNTN71 <- ALNTN67	8	0.567	0.500	0.634
ALNTN71 <- SES	9	-0.207	-0.281	-0.133

### ML Estimates of Free Parameters in Dependence Relationships (contd...)

Path	Standard Error	t
ANOMIA67 <- ALNTN67	0.025	30.732
POWRLS67 <- ALNTN67	0.026	33.064
ANOMIA71 <- ALNTN71	0.026	31.026
POWRLS71 <- ALNTN71	0.027	31.188
EDUCTN <- SES	0.032	26.439
SEI <- SES	0.030	21.297
ALNTN67 <- SES	0.035	-16.263
ALNTN71 <- ALNTN67	0.041	13.880
ALNTN71 <- SES	0.045	-4.603

### Scaled Standard Deviation (nuisance parameters)

Variable	Estimate
ANOMIA67	1.000
POWRLS67	1.000
ANOMIA71	1.000
POWRLS71	1.000
EDUCTN	1.000
SEI	1.000

### Values of Fixed Parameters in Dependence Relationships

Path	Value
ANOMIA67 <- E1	1.000
POWRLS67 <- E2	1.000
ANOMIA71 <- E3	1.000
POWRLS71 <- E4	1.000
EDUCTN <- D1	1.000
SEI <- D2	1.000
ALNTN67 <- Z1	1.000
ALNTN71 <- Z2	1.000

### ML Estimates of Free Parameters in Variance/Covariance Relationships

Path	Parameter Number	Point Estimate	90.00% Confidence Interval		Standard Error
			Lower	Upper	
E1 <-> E1	10	0.400	0.341	0.470	0.039
E1 <-> E3	11	0.133	0.091	0.175	0.026
E2 <-> E2	12	0.274	0.211	0.357	0.044
E2 <-> E4	13	0.035	-0.009	0.080	0.027
E3 <-> E3	14	0.351	0.289	0.427	0.042
E4 <-> E4	15	0.308	0.243	0.390	0.044
D1 <-> D1	16	0.292	0.216	0.395	0.054
D2 <-> D2	17	0.588	0.528	0.656	0.039
Z1 <-> Z1	18	0.683	0.616	0.743	0.039
Z2 <-> Z2	19	0.503	0.448	0.557	0.033



## ML Estimates of Free Parameters in Variance/Covariance Relationships (contd...)

Path	t
E1 <-> E1	10.252
E1 <-> E3	5.216
E2 <-> E2	6.241
E2 <-> E4	1.299
E3 <-> E3	8.400
E4 <-> E4	6.936
D1 <-> D1	5.443
D2 <-> D2	15.219
Z1 <-> Z1	17.518
Z2 <-> Z2	15.084

## Values of Fixed Parameters in Variance/Covariance Relationships

Path	Value
SES <-> SES	1.000

## Equality Constraints on Variances

Constraint	Value	Lagrange Multiplier	Standard Error
ALNTN71 <-> ALNTN71	1.000	0.000	0.000
ALNTN67 <-> ALNTN67	1.000	0.000	0.000
ANOMIA67 <-> ANOMIA67	1.000	0.000	0.000
POWRLS67 <-> POWRLS67	1.000	0.000	0.000
ANOMIA71 <-> ANOMIA71	1.000	0.000	0.000
POWRLS71 <-> POWRLS71	1.000	0.000	0.000
EDUCTN <-> EDUCTN	1.000	0.000	0.000
SEI <-> SEI	1.000	0.000	0.000

## Maximum Likelihood Discrepancy Function

## Measures of Fit of the Model

Sample Discrepancy Function Value : 0.005 (0.005)

## Population Discrepancy Function Value, Fo

Bias Adjusted Point Estimate : 0.001  
90% Confidence Interval : (0.000, 0.011)

## Root Mean Square Error of Approximation (RMSEA)

Steiger-Lind : RMSEA =  $\sqrt{Fo/df}$  : 0.014  
Point Estimate (modified AIC) : (0.000, 0.053)  
90% Confidence Interval

## Expected Cross-Validation Index (CVI)

Point Estimate (modified AIC) : 0.042  
90% Confidence Interval : (0.041, 0.052)  
CVI (modified AIC) for the Saturated Model : 0.045  
Test Statistic : 4.739

Exceedance Probabilities : 0.315  
Ho: Perfect Fit (RMSEA = 0.0) : 0.929  
Ho: Close Fit (RMSEA  $\leq$  0.050)

Multiplier for Obtaining Test Statistic : 931.000  
Degrees of Freedom : 4  
Effective Number of Parameters : 17

After a summary of the input specifications, SYSTAT produces details of the iteration process. The number of the step-halving step, carried out to yield a reduction in the discrepancy function plus a penalty for constraint violations, is given in parentheses next to the iteration number. *Method* indicates the method of estimation. *Discr. Funct.* reports the discrepancy function value. *Max. R. Cos.* equals the absolute value of the maximum residual cosine used to indicate convergence. *Max. Const.* is the absolute value of the maximum violated variance constraint. This panel also includes the number of apparently redundant parameters (number of zero pivots of the coefficient matrix of the normal equations—*NRP*) and the number of active bounds on parameter values (*NBD*).

The values of *NRP* and *NBD* can change from iteration to iteration. If *NRP* has a constant nonzero value for several iterations prior to convergence, this suggests that the model could be overparameterized. The value of *NBD* indicates the number of variance or correlation estimates on bounds at any iteration.

Next, the output includes three matrices: the sample correlation (covariance) matrix, the correlation (covariance) matrix reproduced by the model, and the matrix of residuals. The residual matrix is the difference between the sample correlation (covariance) matrix and the reproduced correlation (covariance) matrix. If the input is a correlation matrix (*TYPE = CORR*), the residual matrix will have null diagonal elements.

For both the dependence and covariance relationships, SYSTAT prints estimates of the free-path coefficients and the values of all fixed-path coefficients involved in the model. The following values are reported for the free parameters:

- *Path.*
- *Param #.* The number of the parameter. This number need not be the same as the number in the input file. (It is the number assigned to the parameter name in the asymptotic covariance matrix of estimators given subsequently.)
- *Point Estimate.* The estimate of the path coefficient.
- *90.00% Conf. Int.* A 90% confidence interval for the path coefficient (the default). If you want to alter the confidence level, specify, for example, *CONF1 = 0.95*.
- *Standard Error.* An estimate of the standard error of the estimator.
- *T value.* The value of the *t* statistic (ratio of estimate to standard error).

If the input is a correlation matrix, the scaled standard deviations (nuisance parameters) are reported with:

- The name of the manifest variable.

- The ratio of the standard deviation reproduced from the model to the sample standard deviation.

After the covariance relationship output, SYSTAT presents information about equality constraints on endogenous variable variances (if applicable):

- *Constraint*. The variance path that is constrained.
- *Value*. The value of the endogenous variable variance at convergence.
- *Lagrange Multiplier*. The value of the Lagrange multiplier at convergence.
- *Standard Error*. An estimate of the standard error of the Lagrange multiplier.

In most applications, the constraints on endogenous variable variances serve as identification conditions and all Lagrange multipliers and standard errors are 0.

## Example 2

### Path Analysis with a Restart File

This example is based on Jöreskog's (1977) path analysis model for the Duncan, Haller, and Portes (1971) data on peer influences on ambition. It illustrates a situation where some manifest variables are exogenous. It also illustrates the use of a restart file for creating a data file for a second run where some modifications have been made.

The example consists of two runs. Jöreskog's original model is used for the first run. The model is treated as a covariance structure—this is inappropriate because a correlation matrix is used as input. In the second run, we use a restart file that treats the model as a correlation structure.

The six manifest exogenous variables are:

<i>RPARASP</i>	Respondent's parental aspiration
<i>RESOCIEC</i>	Respondent's socioeconomic status
<i>REINTGCE</i>	Respondent's intelligence
<i>BFINTGCE</i>	Best friend's intelligence
<i>BFSOCIEC</i>	Best friend's socioeconomic status
<i>BFPARASP</i>	Best friend's parental aspiration

The four endogenous variables are:

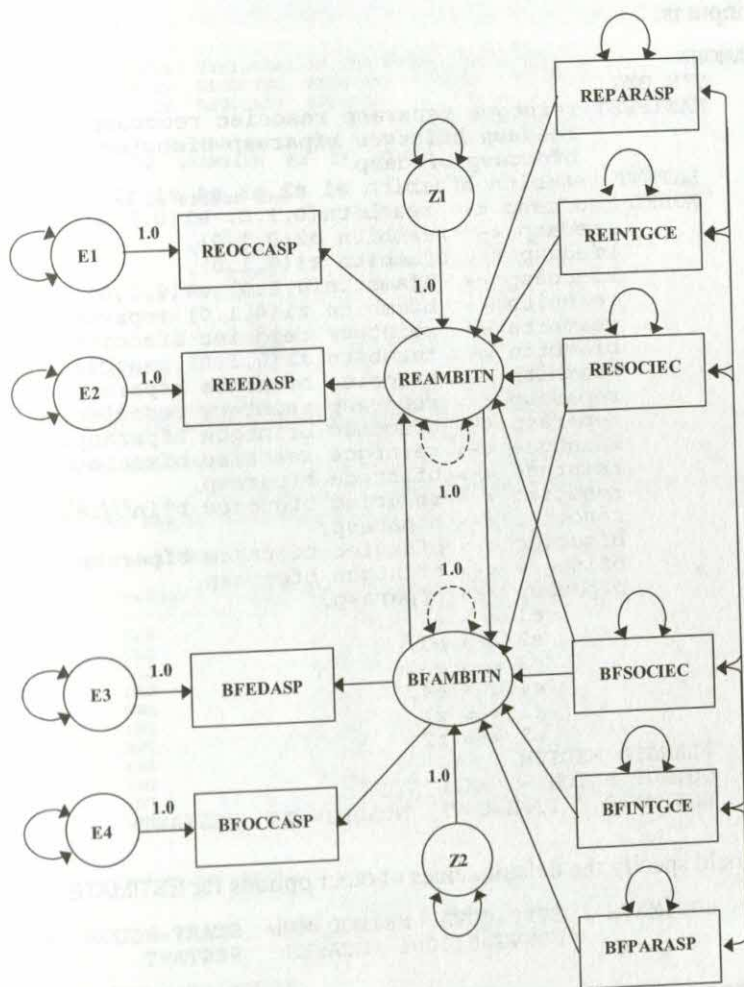
<i>REOCCASP</i>	Respondent's occupational aspiration
<i>BFEDASP</i>	Best friend's educational aspiration
<i>REEDASP</i>	Respondent's educational aspiration
<i>BFOCCASP</i>	Best friend's occupational aspiration

The latent endogenous variables are:

<i>REAMBITN</i>	Respondent's ambition
<i>BFAMBITN</i>	Best friend's ambition

And the exogenous error variables are *E1*, *E4*, *E2*, *Z1*, *E3*, and *Z2*.





The correlation matrix for the manifest variables is stored in the file *EX2*.

The input is:

```

RAMONA
USE EX2
MANIFEST reintgce reparasp resociec reoccasp,
         reedasp bfintgce bfparasp bfsociec,
         bfoccasp bfedasp
LATENT reambitn bfambitn e1 e2 e3 e4 z1 z2
MODEL reoccasp <- reambitn(0,1.0) e1(0,1.0),
      reedasp <- reambitn e2(0,1.0),
      bfedasp <- bfambitn e3(0,1.0),
      bfoccasp <- bfambitn(0,1.0) e4(0,1.0),
      reambitn <- bfambitn z1(0,1.0) reparasp,
      reambitn <- reintgce resociec bfsociec,
      bfambitn <- reambitn z2(0,1.0) resociec,
      bfambitn <- bfsociec bfintgce bfparasp,
      reparasp <-> reparasp reintgce resociec,
      reparasp <-> bfsociec bfintgce bfparasp,
      reintgce <-> reintgce resociec bfsociec,
      reintgce <-> bfintgce bfparasp,
      resociec <-> resociec bfsociec bfintgce,
      resociec <-> bfparasp,
      bfsociec <-> bfsociec bfintgce bfparasp,
      bfintgce <-> bfintgce bfparasp,
      bfparasp <-> bfparasp,
      e1 <-> e1,
      e2 <-> e2,
      e3 <-> e3,
      e4 <-> e4,
      z1 <-> z1,
      z2 <-> z2
PLENGTH MEDIUM
OUTPUT BATCH = 'EX2B.SYC'
ESTIMATE / TYPE=COVA NCASES=329 RESTART

```

You would specify the default values of other options for ESTIMATE as:

```

ESTIMATE / TYPE=COVA METHOD=MWL START=ROUGH ITER=500,
          CONVG=0.0001 NCASES RESTART

```

The RESTART option of ESTIMATE creates a restart command file, *EX2B.SYC*, that is submitted as the input in the second run. RESTART tells RAMONA to take the estimated parameter values and insert them as starting values in the MODEL statement. Note that we must also type OUTPUT BATCH = *filename* to do this. Before the second run, we modify *EX2B.SYC* to treat the model as a correlation structure.

Following Jöreskog's model, the path coefficients REOCCASP <- REAMBITN and BFOCCASP <- BFAMBITN are set equal to 1 for identification purposes.

The output is:

There are 10 Manifest Variables in the Model. They are  
 REINTGCE REPARASP RESOCIEC REOCCASP REEDASP BFINTGCE  
 BFPARASP BFSOCIEC BFOCCASP BFEDASP

There are 8 Latent Variables in the Model. They are  
 REAMBITN E1 E2 BFAMBITN E3 E4 Z1 Z2

RAMONA Options in Effect are

Display	:	Covar
Method	:	MWL
Start	:	Rough
Convergence Limit	:	0.0001
Maximum Iterations	:	100
N of Cases	:	329
Restart	:	Yes
% Confidence Level	:	90

Number of Manifest Variables : 10  
 Total Number of Variables in the System : 18

Reading Correlation Matrix...

\*\*\* WARNING \*\*\* : A correlation matrix was provided although DISP=COV fit measures and standard errors may be inappropriate.

#### Details of Iterations

Iteration	Method	Discr. Funct.	Max.R.Cos.	Max.Const.	NRP	NBD
	OLS	1.501			0	0
1(0)	OLS	0.325	0.720		0	0
2(0)	OLS	0.023	0.191		0	0
3(0)	OLS	0.020	0.007		0	0
3(0)	MWL	0.085	0.060		0	0
4(0)	MWL	0.082	0.017		0	0
5(0)	MWL	0.082	0.004		0	0
6(0)	MWL	0.082	0.001		0	0
7(0)	MWL	0.082	0.000		0	0
8(0)	MWL	0.082	0.000		0	0
9(0)	MWL	0.082	0.000		0	0

Iterative procedure complete.

Convergence Limit for Residual Cosines: 1.000E-04 on 2 Consecutive Iterations

#### Sample Covariance Matrix

	REINTGCE	REPARASP	RESOCIEC	REOCCASP	REEDASP	BFINTGCE	BFPARASP
REINTGCE	1.000						
REPARASP	0.184	1.000					
RESOCIEC	0.222	0.049	1.000				
REOCCASP	0.410	0.214	0.324	1.000			
REEDASP	0.404	0.274	0.405	0.625	1.000		
BFINTGCE	0.336	0.078	0.230	0.299	0.286	1.000	
BFPARASP	0.102	0.115	0.093	0.076	0.070	0.209	1.000
BFSOCIEC	0.186	0.019	0.271	0.293	0.241	0.295	-0.044
BFOCCASP	0.260	0.084	0.279	0.422	0.328	0.501	0.199
BFEDASP	0.290	0.112	0.305	0.327	0.367	0.519	0.278

## Sample Covariance Matrix (contd...)

	BFSOCIEC	BFOCCASP	BFEDASP
REINTGCE			
REPARASP			
RESOCIEC			
REOCCASP			
REEDASP			
BFINTGCE			
BFPARASP			
BFSOCIEC	1.000		
BFOCCASP	0.361	1.000	
BFEDASP	0.410	0.640	1.000

Number of Cases : 329

## Reproduced Covariance Matrix

	REINTGCE	REPARASP	RESOCIEC	REOCCASP	REEDASP	BFINTGCE	BFPARASP
REINTGCE	1.000						
REPARASP	0.184	1.000					
RESOCIEC	0.222	0.049	1.000				
REOCCASP	0.393	0.239	0.357	0.999			
REEDASP	0.417	0.254	0.379	0.623	0.999		
BFINTGCE	0.336	0.078	0.230	0.258	0.274	1.000	
BFPARASP	0.102	0.115	0.093	0.103	0.110	0.209	1.000
BFSOCIEC	0.186	0.019	0.271	0.255	0.270	0.295	-0.044
BFOCCASP	0.255	0.095	0.282	0.330	0.351	0.489	0.237
BFEDASP	0.273	0.102	0.303	0.354	0.376	0.525	0.254

## Reproduced Covariance Matrix (contd...)

	BFSOCIEC	BFOCCASP	BFEDASP
REINTGCE			
REPARASP			
RESOCIEC			
REOCCASP			
REEDASP			
BFINTGCE			
BFPARASP			
BFSOCIEC	1.000		
BFOCCASP	0.374	0.999	
BFEDASP	0.401	0.639	0.999

## Residual Matrix (covariances)

	REINTGCE	REPARASP	RESOCIEC	REOCCASP	REEDASP	BFINTGCE	BFPARASP
REINTGCE	0.000						
REPARASP	0.000	0.000					
RESOCIEC	0.000	0.000	0.000				
REOCCASP	0.018	-0.026	-0.033	0.001			
REEDASP	-0.013	0.020	0.026	0.001	0.001		
BFINTGCE	0.000	0.000	0.000	0.042	0.013	0.000	
BFPARASP	0.000	0.000	0.000	-0.027	-0.039	0.000	0.000
BFSOCIEC	0.000	0.000	0.000	0.038	-0.030	0.000	0.000
BFOCCASP	0.005	-0.011	-0.004	0.091	-0.023	0.011	-0.038
BFEDASP	0.017	0.010	0.003	-0.027	-0.009	-0.006	0.024



## Residual Matrix (covariances) (contd...)

	BFSOCIEC	BFOCCASP	BFEDASP
REINTGCE			
REPARASP			
RESOCIEC			
REOCCASP			
REEDASP			
BFINTGCE			
BFPARASP			
BFSOCIEC	0.000		
BFOCCASP	-0.013	0.001	
BFEDASP	0.009	0.001	0.001

Value of the Maximum Absolute Residual : 0.091

## ML Estimates of Free Parameters in Dependence Relationships

Path	Parameter Number	Point Estimate	90.00% Confidence Interval	
			Lower	Upper
REEDASP <- REAMBITN	1	1.062	0.914	1.210
BFEDASP <- BFAMBITN	2	1.073	0.940	1.206
REAMBITN <- BFAMBITN	3	0.174	0.032	0.316
REAMBITN <- REPARASP	4	0.164	0.100	0.228
REAMBITN <- REINTGCE	5	0.255	0.185	0.324
REAMBITN <- RESOCIEC	6	0.222	0.151	0.294
REAMBITN <- BFSOCIEC	7	0.079	0.001	0.156
BFAMBITN <- REAMBITN	8	0.185	0.054	0.317
BFAMBITN <- RESOCIEC	9	0.067	-0.004	0.138
BFAMBITN <- BFSOCIEC	10	0.218	0.151	0.284
BFAMBITN <- BFINTGCE	11	0.330	0.262	0.398
BFAMBITN <- BFPARASP	12	0.152	0.092	0.212

## ML Estimates of Free Parameters in Dependence Relationships (contd...)

Path	Standard Error	t
REEDASP <- REAMBITN	0.090	11.804
BFEDASP <- BFAMBITN	0.081	13.233
REAMBITN <- BFAMBITN	0.086	2.022
REAMBITN <- REPARASP	0.039	4.228
REAMBITN <- REINTGCE	0.043	5.988
REAMBITN <- RESOCIEC	0.043	5.111
REAMBITN <- BFSOCIEC	0.047	1.675
BFAMBITN <- REAMBITN	0.080	2.327
BFAMBITN <- RESOCIEC	0.043	1.546
BFAMBITN <- BFSOCIEC	0.040	5.379
BFAMBITN <- BFINTGCE	0.041	7.974
BFAMBITN <- BFPARASP	0.036	4.181

## Values of Fixed Parameters in Dependence Relationships

Path	Value
REOCCASP <- REAMBITN	1.000
REOCCASP <- E1	1.000
REEDASP <- E2	1.000
BFEDASP <- E3	1.000
BFOCCASP <- BFAMBITN	1.000
BFOCCASP <- E4	1.000
REAMBITN <- Z1	1.000
BFAMBITN <- Z2	1.000

## ML Estimates of Free Parameters in Variance/Covariance Relationships

Path	Parameter Number	Point Estimate	90.00% Confidence Interval	
			Lower	Upper
REPARASP <-> REPARASP	13	1.000	0.879	1.137
REPARASP <-> REINTGCE	14	0.184	0.092	0.276
REPARASP <-> RESOCIEC	15	0.049	-0.042	0.140
REPARASP <-> BFSOCIEC	16	0.019	-0.072	0.109
REPARASP <-> BFINTGCE	17	0.078	-0.013	0.169
REPARASP <-> BFPARASP	18	0.115	0.023	0.206
REINTGCE <-> REINTGCE	19	1.000	0.879	1.137
REINTGCE <-> RESOCIEC	20	0.222	0.129	0.315
REINTGCE <-> BFSOCIEC	21	0.186	0.094	0.278
REINTGCE <-> BFINTGCE	22	0.336	0.240	0.431
REINTGCE <-> BFPARASP	23	0.102	0.011	0.193
RESOCIEC <-> RESOCIEC	24	1.000	0.879	1.137
RESOCIEC <-> BFSOCIEC	25	0.271	0.177	0.365
RESOCIEC <-> BFINTGCE	26	0.230	0.137	0.323
RESOCIEC <-> BFPARASP	27	0.093	0.002	0.184
BFSOCIEC <-> BFSOCIEC	28	1.000	0.879	1.137
BFSOCIEC <-> BFINTGCE	29	0.295	0.200	0.390
BFSOCIEC <-> BFPARASP	30	-0.044	-0.135	0.047
BFINTGCE <-> BFINTGCE	31	1.000	0.879	1.137
BFINTGCE <-> BFPARASP	32	0.209	0.116	0.301
BFPARASP <-> BFPARASP	33	1.000	0.879	1.137
E1 <-> E1	34	0.412	0.336	0.506
E2 <-> E2	35	0.337	0.262	0.434
E3 <-> E3	36	0.313	0.246	0.399
E4 <-> E4	37	0.404	0.335	0.487
Z1 <-> Z1	38	0.281	0.214	0.370
Z2 <-> Z2	39	0.229	0.173	0.303

## ML Estimates of Free Parameters in Variance/Covariance Relationships (contd...)

Path	Standard Error	t
REPARASP <-> REPARASP	0.078	12.806
REPARASP <-> REINTGCE	0.056	3.276
REPARASP <-> RESOCIEC	0.055	0.885
REPARASP <-> BFSOCIEC	0.055	0.337
REPARASP <-> BFINTGCE	0.055	1.412
REPARASP <-> BFPARASP	0.056	2.064
REINTGCE <-> REINTGCE	0.078	12.806
REINTGCE <-> RESOCIEC	0.057	3.925
REINTGCE <-> BFSOCIEC	0.056	3.314
REINTGCE <-> BFINTGCE	0.058	5.761
REINTGCE <-> BFPARASP	0.056	1.840
RESOCIEC <-> RESOCIEC	0.078	12.806
RESOCIEC <-> BFSOCIEC	0.057	4.732
RESOCIEC <-> BFINTGCE	0.057	4.063
RESOCIEC <-> BFPARASP	0.055	1.679
BFSOCIEC <-> BFSOCIEC	0.078	12.806
BFSOCIEC <-> BFINTGCE	0.058	5.124
BFSOCIEC <-> BFPARASP	0.055	-0.792
BFINTGCE <-> BFINTGCE	0.078	12.806
BFINTGCE <-> BFPARASP	0.056	3.700
BFPARASP <-> BFPARASP	0.078	12.806
E1 <-> E1	0.051	8.068
E2 <-> E2	0.052	6.503
E3 <-> E3	0.046	6.844
E4 <-> E4	0.046	8.754
Z1 <-> Z1	0.047	6.029
Z2 <-> Z2	0.039	5.859

**Maximum Likelihood Discrepancy Function****Measures of Fit of the Model**

Sample Discrepancy Function Value : 0.082 (0.082)

**Population Discrepancy Function Value,  $F_0$** 

Bias Adjusted Point Estimate : 0.033  
90% Confidence Interval : (0.001, 0.089)

**Root Mean Square Error of Approximation (RMSEA)**

Steiger-Lind :  $RMSEA = \sqrt{F_0/df}$   
Point Estimate (modified AIC) : 0.046  
90% Confidence Interval : (0.008, 0.075)

**Expected Cross-Validation Index (CVI)**

Point Estimate (modified AIC) : 0.320  
90% Confidence Interval : (0.288, 0.376)  
CVI (modified AIC) for the Saturated Model : 0.335

Test Statistic : 26.893

Exceedance Probabilities  
Ho: Perfect Fit ( $RMSEA = 0.0$ ) : 0.043  
Ho: Close Fit ( $RMSEA \leq 0.050$ ) : 0.560

Multiplier for Obtaining Test Statistic : 328.000  
Degrees of Freedom : 16  
Effective Number of Parameters : 39

**Using the Restart File**

A restart file was created during the first run to form an input file that specifies the model represented in the path diagram. Now type the following modifications into the *EX2B* restart file and save the file:

- $DISP = COVA$  is replaced by  $DISP = CORR$ .
- $START = ROUGH$  is replaced by  $START = CLOSE$ .
- $REOCCASP <- REAMBITN(0,1.0)$  is replaced by  $REOCCASP <- REAMBITN(*,1.0)$ , freeing a fixed-path coefficient.
- $BFOCCASP <- BFAMBITN(0,1.0)$  is replaced by  $BFOCCASP <- BFAMBITN(*,1.0)$ , freeing a fixed-path coefficient.
- $REAMBITN <-> REAMBITN(0,1.0)$  is added, imposing a variance constraint on an endogenous latent variable.
- $BFAMBITN <-> BFAMBITN(0,1.0)$  is added, imposing a variance constraint on an endogenous latent variable.
- The output is displayed for  $PLENGTH MEDIUM$ .

The modified restart file is shown below:

RAMONA

USE EX2

```
MODEL reoccasp <- reambitn(*,1.000),
      reoccasp <- e1(0,1.000),
      reedasp <- reambitn(1,1.062),
      reedasp <- e2(0,1.000),
      bfedasp <- bfambitn(2,1.073),
      bfedasp <- e3(0,1.000),
      bfoccasp <- bfambitn(*,1.000),
      bfoccasp <- e4(0,1.000),
      reambitn <- bfambitn(3,0.174),
      reambitn <- z1(0,1.000),
      reambitn <- reparasp(4,0.164),
      reambitn <- reintgce(5,0.255),
      reambitn <- resociec(6,0.222),
      reambitn <- bfsociec(7,0.079),
      bfambitn <- reambitn(8,0.185),
      bfambitn <- z2(0,1.000),
      bfambitn <- resociec(9,0.067),
      bfambitn <- bfsociec(10,0.218),
      bfambitn <- bfintgce(11,0.330),
      bfambitn <- bfparasp(12,0.152),
      reparasp <-> reparasp(13,1.000),
      reparasp <-> reintgce(14,0.184),
      reparasp <-> resociec(15,0.049),
      reparasp <-> bfsociec(16,0.019),
      reparasp <-> bfintgce(17,0.078),
      reparasp <-> bfparasp(18,0.115),
      reintgce <-> reintgce(19,1.000),
      reintgce <-> resociec(20,0.222),
      reintgce <-> bfsociec(21,0.186),
      reintgce <-> bfintgce(22,0.336),
      reintgce <-> bfparasp(23,0.102),
      resociec <-> resociec(24,1.000),
      resociec <-> bfsociec(25,0.271),
      resociec <-> bfintgce(26,0.230),
      resociec <-> bfparasp(27,0.093),
      bfsociec <-> bfsociec(28,1.000),
      bfsociec <-> bfintgce(29,0.295),
      bfsociec <-> bfparasp(30,-0.044),
      bfintgce <-> bfintgce(31,1.000),
      bfintgce <-> bfparasp(32,0.209),
      bfparasp <-> bfparasp(33,1.000),
      e1 <-> e1(34,0.412),
```



```

e4      <-> e4 (37,
z1      <-> z1 (38,
z2      <-> z2 (39,
reambitn <-> reambi
bfambitn <-> bfambi
PLENGTH MEDIUM
ESTIMATE / CONVG =0.0001,
METHOD =MWL, ST

```

Note that we rounded some parameters to the START setting, ROUGH, has been chosen, and a restart file is used.

Now execute this modified file in the Commandspace or another terminal.

The input is:

```
SUBMIT EX2B
```

The output is:

```

There are 10 Manifest Variables in the Model
REINTGCE REPARASP RESOCIEC REOCCAS
BFPARASP BFSOCIEC BFOCCASP BFED

```

```

There are 8 Latent Variables in the Model
REAMBITN E1 E2 BFAMBITN E3 E4

```

#### RAMONA Options in Effect are

Display	Corr
Method	MWL
Start	Close
Convergence Limit	0.0001
Maximum Iterations	100
N of Cases	329
Restart	No
% Confidence Level	90

```

Number of Manifest Variables
Total Number of Variables in the System

```

Reading Correlation Matrix...

#### Details of Iterations

Iteration	Method	Discr. Funct.
0	MWL	0.082
1 (0)	MWL	0.082
2 (0)	MWL	0.082
3 (0)	MWL	0.082

```

0.404),
0.281),
0.229),
tn(0,1.000),
tn(0,1.000)

```

```

MAXIT =100, RELKURT =1.000, DISP =CORR,
TART =CLOSE, NCASES =329, CONFI=0.900

```

er values to shorten the commands. Also, the  
 anged to CLOSE (under ESTIMATE) because a

e (after you have edited it and saved it using  
 text editor.

```

he Model. They are
CASP REEDASP BFINTGCE
ASP

```

```

Model. They are
Z1 Z2

```

```

: 10
tem : 28

```

Max.R.Cos.	Max.Const.	NRP	NBD
	0.000		
0.000	0.000	0	0
0.000	0.000	0	0
0.000	0.000	0	0

### Sample Correlation Matrix

	REINTGCE	REPARASP	RESOCIEC
REINTGCE	1.000		
REPARASP	0.184	1.000	
RESOCIEC	0.222	0.049	1.000
REOCCASP	0.410	0.214	0.324
REEDASP	0.404	0.274	0.405
BFINTGCE	0.336	0.078	0.230
BFPARASP	0.102	0.115	0.093
BFSOCIEC	0.186	0.019	0.271
BFOCCASP	0.260	0.084	0.279
BFEDASP	0.290	0.112	0.305

### Sample Correlation Matrix (contd...)

	BFSOCIEC	BFOCCASP	BFEDASP
REINTGCE			
REPARASP			
RESOCIEC			
REOCCASP			
REEDASP			
BFINTGCE			
BFPARASP			
BFSOCIEC	1.000		
BFOCCASP	0.361	1.000	
BFEDASP	0.410	0.640	1.000

Number of Cases : 329

### Reproduced Correlation Matrix

	REINTGCE	REPARASP	RESOCIEC
REINTGCE	1.000		
REPARASP	0.184	1.000	
RESOCIEC	0.222	0.049	1.000
REOCCASP	0.393	0.240	0.357
REEDASP	0.417	0.254	0.379
BFINTGCE	0.336	0.078	0.230
BFPARASP	0.102	0.115	0.093
BFSOCIEC	0.186	0.019	0.271
BFOCCASP	0.255	0.095	0.282
BFEDASP	0.273	0.102	0.303

### Reproduced Correlation Matrix (contd...)

	BFSOCIEC	BFOCCASP	BFEDASP
REINTGCE			
REPARASP			
RESOCIEC			
REOCCASP			
REEDASP			
BFINTGCE			
BFPARASP			
BFSOCIEC	1.000		

REOCCASP	REEDASP	BFINTGCE	BFPARASP
----------	---------	----------	----------

---

1.000			
0.625	1.000		
0.299	0.286	1.000	
0.076	0.070	0.209	1.000
0.293	0.241	0.295	-0.044
0.422	0.328	0.501	0.199
0.327	0.367	0.519	0.278

REOCCASP	REEDASP	BFINTGCE	BFPARASP
----------	---------	----------	----------

---

1.000			
0.624	1.000		
0.258	0.274	1.000	
0.103	0.110	0.209	1.000
0.255	0.270	0.295	-0.044
0.330	0.351	0.489	0.237
0.355	0.376	0.525	0.254



	REINTGCE	REPARASP	R
REINTGCE	0.000		
REPARASP	0.000	0.000	
RESOCIEC	0.000	0.000	
REOCCASP	0.017	-0.026	
REEDASP	-0.013	0.020	
BFINTGCE	0.000	0.000	
BFPARASP	0.000	0.000	
BFSOCIEC	0.000	0.000	
BFOCCASP	0.005	-0.011	
BFEDASP	0.017	0.010	
<b>Residual Matrix (correlations)</b>			

	BFSOCIEC	BFOCCASP	B
REINTGCE			
REPARASP			
RESOCIEC			
REOCCASP			
REEDASP			
BFINTGCE			
BFPARASP			
BFSOCIEC	0.000		
BFOCCASP	-0.013	0.000	
BFEDASP	0.009	0.001	

Value of the Maximum Absolute Residual

#### ML Estimates of Free Parameters in

Path	Parameter Number
REOCCASP <- REAMBITN	1
REEDASP <- REAMBITN	2
BFEDASP <- BFAMBITN	3
BFOCCASP <- BFAMBITN	4
REAMBITN <- BFAMBITN	5
REAMBITN <- REPARASP	6
REAMBITN <- REINTGCE	7
REAMBITN <- RESOCIEC	8
REAMBITN <- BFSOCIEC	9
BFAMBITN <- REAMBITN	10
BFAMBITN <- RESOCIEC	11
BFAMBITN <- BFSOCIEC	12
BFAMBITN <- BFINTGCE	13
BFAMBITN <- BFPARASP	14

#### ML Estimates of Free Parameters in

Path	Standard Error
REOCCASP <- REAMBITN	0.000
REEDASP <- REAMBITN	0.000
BFEDASP <- BFAMBITN	0.000
BFOCCASP <- BFAMBITN	0.000
REAMBITN <- BFAMBITN	0.000
REAMBITN <- REPARASP	0.000
REAMBITN <- REINTGCE	0.000
REAMBITN <- RESOCIEC	0.000
REAMBITN <- BFSOCIEC	0.000
BFAMBITN <- REAMBITN	0.000

ESOCIEC	REOCCASP	REEDDSP	BFINTGCE	BFPARASP
0.000				
-0.033	0.000			
0.025	0.001	0.000		
0.000	0.042	0.012	0.000	
0.000	-0.027	-0.039	0.000	0.000
0.000	0.038	-0.030	0.000	0.000
-0.004	0.091	-0.023	0.011	-0.038
0.002	-0.028	-0.010	-0.006	0.024

contd...)

BFEDASP

0.000

nal : 0.091

# Dependence Relationships

Point Estimate	90.00% Confidence Interval	
	Lower	Upper
0.766	0.710	0.823
0.814	0.759	0.868
0.828	0.781	0.876
0.772	0.721	0.823
0.175	0.034	0.317
0.214	0.133	0.294
0.332	0.248	0.417
0.290	0.201	0.378
0.103	0.002	0.204
0.184	0.055	0.313
0.087	-0.005	0.178
0.282	0.200	0.365
0.428	0.349	0.506
0.197	0.121	0.273

# Dependence Relationships (contd...)

error t

0.034	22.215
0.033	24.523
0.029	28.486
0.031	24.748
0.086	2.036
0.049	4.363
0.051	6.465
0.054	5.386
0.061	1.685
0.078	2.350

REOCCASP	1.000
REEDASP	1.000
BFOCCASP	1.000
BFEDASP	1.000
REPARASP	1.000
BFINTGCE	1.000
BFPARASP	1.000
BFSOCIEC	1.000
RESOCIEC	1.000
REINTGCE	1.000

# Values of Fixed Parameters in Dependence R

Path	Value
REOCCASP <- E1	1.000
REEDASP <- E2	1.000
BFEDASP <- E3	1.000
BFOCCASP <- E4	1.000
REAMBITN <- Z1	1.000
BFAMBITN <- Z2	1.000

# ML Estimates of Free Parameters in Variance

Path	Parameter Number	Point Es
REPARASP <-> REINTGCE	15	
REPARASP <-> RESOCIEC	16	
REPARASP <-> BFSOCIEC	17	
REPARASP <-> BFINTGCE	18	
REPARASP <-> BFPARASP	19	
REINTGCE <-> RESOCIEC	20	
REINTGCE <-> BFSOCIEC	21	
REINTGCE <-> BFINTGCE	22	
REINTGCE <-> BFPARASP	23	
RESOCIEC <-> BFSOCIEC	24	
RESOCIEC <-> BFINTGCE	25	
RESOCIEC <-> BFPARASP	26	
BFSOCIEC <-> BFINTGCE	27	
BFSOCIEC <-> BFPARASP	28	
BFINTGCE <-> BFPARASP	29	
E1 <-> E1	30	
E2 <-> E2	31	
E3 <-> E3	32	
E4 <-> E4	33	
Z1 <-> Z1	34	
Z2 <-> Z2	35	

# ML Estimates of Free Parameters in Variance/

Path	Standard Error	
REPARASP <-> REINTGCE	0.053	3.
REPARASP <-> RESOCIEC	0.055	0.
REPARASP <-> BFSOCIEC	0.055	0.
REPARASP <-> BFINTGCE	0.055	1.
REPARASP <-> BFPARASP	0.054	2.
REINTGCE <-> RESOCIEC	0.052	4.
REINTGCE <-> BFSOCIEC	0.053	3.
REINTGCE <-> BFINTGCE	0.049	

# relationships

## /Covariance Relationships

Estimate	90.00% Confidence Interval	
	Lower	Upper
0.184	0.095	0.270
0.049	-0.042	0.139
0.019	-0.072	0.109
0.078	-0.012	0.168
0.115	0.024	0.203
0.222	0.134	0.306
0.186	0.097	0.272
0.336	0.253	0.414
0.102	0.012	0.191
0.271	0.185	0.353
0.230	0.143	0.314
0.093	0.003	0.182
0.295	0.210	0.376
-0.044	-0.134	0.047
0.209	0.120	0.294
0.413	0.334	0.509
0.338	0.259	0.439
0.314	0.244	0.404
0.404	0.332	0.492
0.479	0.390	0.570
0.384	0.305	0.470

## Covariance Relationships (contd...)

t

447  
888  
337  
425  
105  
229  
191



BFINTGCE <-> BFPARASP	0.0
E1 <-> E1	0.0
E2 <-> E2	0.0
E3 <-> E3	0.0
E4 <-> E4	0.0
Z1 <-> Z1	0.0
Z2 <-> Z2	0.0

### Values of Fixed Parameters in Variance-Covariance Matrix

Path	Value
REPARASP <-> REPARASP	1.000
REINTGCE <-> REINTGCE	1.000
RESOCIEC <-> RESOCIEC	1.000
BFSOCIEC <-> BFSOCIEC	1.000
BFINTGCE <-> BFINTGCE	1.000
BFPARASP <-> BFPARASP	1.000

### Equality Constraints on Variances

Constraint	Value	Mu
REAMBITN <-> REAMBITN	1.000	
BFAMBITN <-> BFAMBITN	1.000	
REOCCASP <-> REOCCASP	1.000	
REEDASP <-> REEDASP	1.000	
BFOCCASP <-> BFOCCASP	1.000	
BFEDASP <-> BFEDASP	1.000	

### Maximum Likelihood Discrepancy Function

### Measures of Fit of the Model

Sample Discrepancy Function Value

### Population Discrepancy Function Value

Bias Adjusted Point Estimate  
90% Confidence Interval

### Root Mean Square Error of Approximation

Steiger-Lind :  $RMSEA = \sqrt{Fo/df}$   
Point Estimate (modified AIC)  
90% Confidence Interval

### Expected Cross-Validation Index (CVI)

Point Estimate (modified AIC)  
90% Confidence Interval  
CVI (modified AIC) for the Saturated

Test Statistic

Exceedance Probabilities  
Ho: Perfect Fit ( $RMSEA = 0.0$ )  
Ho: Close Fit ( $RMSEA \leq 0.050$ )

053	3.952
053	7.804
054	6.250
048	6.511
048	8.389
055	8.640
051	7.591

# ance/Covariance Relationships

Lagrange multiplier	Standard Error
------------------------	----------------

0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000

ion

: 0.082 (0.082)

e, Fo

: 0.033  
: (0.001, 0.089)

ion (RMSEA)

: 0.046  
: (0.008, 0.075)

Model

: 0.320  
: (0.288, 0.376)

: 0.335  
: 26.893

: 0.043  
: 0.560

tic : 328.000

runs, but the maximum likelihood estimates conditions. The standard errors in the second (incorrect). An appropriate warning has been last run the Lagrange multipliers and the correlation all equality constraints on endogenous variables conditions, not constraints on the model. The applications.

### ***Example 3***

#### ***Path Analysis Using Rectangular Input***

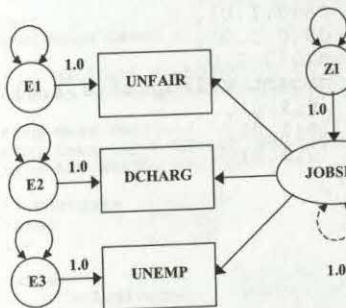
This example (Mels and Koorts, 1989) illustrates a path analysis using a SYSTAT data file. Asymptotically

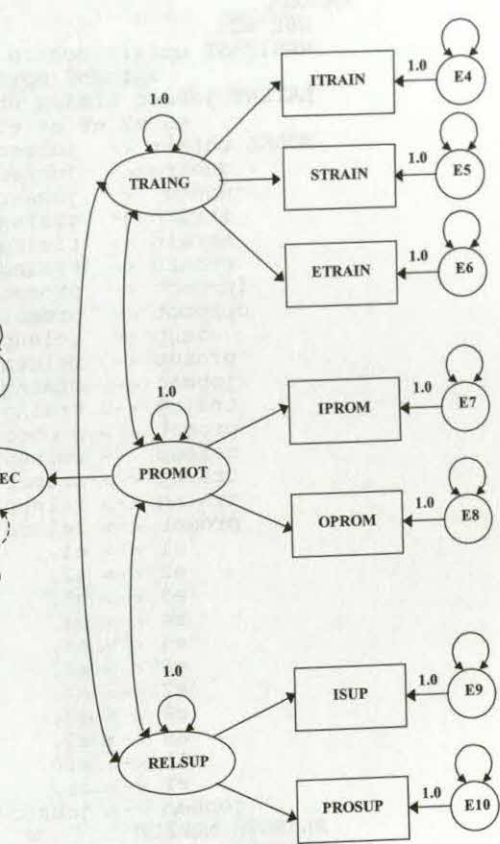
A questionnaire concerned with job satisfaction are 10 manifest variables that serve as indicators (JOBSEC), attitude toward training (TRAINING), promotion (PROMOT), and relations with superiors (RELAT) to account for causal relationships between the

differ because of different identification  
and run differ (those in the first run were  
an output by RAMONA. Notice that in the  
corresponding standard errors are 0 because  
the variances act as identification  
this is the case in most, but not all, practical

estimates how RAMONA uses the usual cases-  
distribution-free estimates are obtained.  
The data collection was completed by 213 nurses. There  
are four latent variables: job security  
(JS), opportunities for promotion  
(OP), and satisfaction (SAT). The path diagram shows a model  
with three latent variables.







RAMONA

USE EX3

MANIFEST unfair dcharg unemp  
                  ipromot opromot i

LATENT jobsec traing promot  
          e6 e7 e8 e9 e10 z1

MODEL unfair <- jobsec e1(0  
          dcharg <- jobsec e2(0  
          unemp <- jobsec e3(0  
          itrain <- traing e4(0  
          strain <- traing e5(0  
          etrain <- traing e6(0  
          ipromot <- promot e7(0  
          opromot <- promot e8(0  
          isup <- relsup e9(0  
          prosup <- relsup e10(0  
          jobsec <- traing prom  
          traing <-> traing (0,1  
          promot <-> promot (0,1  
          relsup <-> relsup (0,1  
          traing <-> promot,  
          traing <-> relsup,  
          promot <-> relsup,  
          e1 <-> e1,  
          e2 <-> e2,  
          e3 <-> e3,  
          e4 <-> e4,  
          e5 <-> e5,  
          e6 <-> e6,  
          e7 <-> e7,  
          e8 <-> e8,  
          e9 <-> e9,  
          e10 <-> e10,  
          z1 <-> z1,  
          jobsec <-> jobsec(0,1.

PLENGTH MEDIUM

ESTIMATE / TYPE=CORR METHOD

itrain strain etrain,  
sup prosup  
relsup e1 e2 e3 e4 e5,

,1.0),  
,1.0),  
,1.0),  
,1.0),  
,1.0),  
,1.0),  
,1.0),  
,1.0),  
,1.0),  
,1.0),  
0,1.0),  
ot relsup z1(0,1.0),  
.0),  
.0),  
.0),

0)

=ADFU



There are 10 Manifest Variables in the  
 UNFAIR DCHARG UNEMP ITRAIN STRA  
 ISUP PROSUP

There are 15 Latent Variables in the  
 JOBSEC E1 E2 E3 TRAING E4 E5  
 E10 Z1

### RAMONA Options in Effect are

Display		Cor
Method		AD
Start		Roug
Convergence Limit		0.000
Maximum Iterations		10
N of Cases	determined whe	data are rea
Restart		
% Confidence Level		

Number of Manifest Variables  
 Total Number of Variables in the Sys

Computing Mean Vector...  
 Computing Covariance Matrix and Four  
 Computing ADF Weight Matrix...

Overall Kurtosis : 19.754  
 Normalized : 9.305  
 Relative : 1.165

Variable	Individual	Kurtosis Normalized
UNFAIR	1.395	4.155
DCHARG	1.866	5.560
UNEMP	0.181	0.540
ITRAIN	-0.560	-1.669
STRAIN	-1.102	-3.282
ETRAIN	-0.730	-2.174
IPROMOT	-1.006	-2.997
OPROMOT	-0.757	-2.250
ISUP	-0.945	-2.811
PROSUP	-0.547	-1.622

Smallest relative pivot of covarian  
 covariances : 0.149

### Details of Iterations

Iteration	Method	Discr. Funct
0	OLS	1.25
1(0)	OLS	0.39
2(0)	OLS	0.07
3(0)	OLS	0.07
4(0)	OLS	0.07
4(0)	ADFU	0.39
5(0)	ADFU	0.19
6(0)	ADFU	0.18
7(0)	ADFU	0.18
8(0)	ADFU	0.18
9(0)	ADFU	0.18
10(0)	ADFU	0.18

Model. They are  
E6 PROMOT E7 E8 RELSUP E9

### th Order Moments...

	Relative
6	1.465
0	1.622
0	1.060
9	0.813
2	0.633
4	0.757
7	0.665
6	0.748
5	0.685
8	0.818

ce matrix of sample

	Max.R.Cos.	Max.Const.	NRP	NBD
5		0.000		
9	0.556	0.405	0	0
9	0.115	0.046	0	0
5	0.011	0.000	0	0
5	0.002	0.000	0	0
3	0.361	0.000	0	0
00	0.085	0.040	0	0
35	0.020	0.005	0	0
35	0.003	0.000	0	0
35	0.002	0.000	0	0
35	0.000	0.000	0	0
35	0.000	0.000	0	0
35	0.000	0.000	0	0
35	0.000	0.000	0	0

**Sample Correlation Matrix**

	UNFAIR	DCHARG	UNEMP	ITRAIN
UNFAIR	1.000			
DCHARG	0.438	1.000		
UNEMP	0.249	0.455	1.000	
ITRAIN	0.150	0.110	0.056	1.000
STRAIN	0.173	0.209	0.028	0.540
ETRAIN	0.184	0.168	-0.006	0.540
IPROMOT	0.134	0.210	0.169	0.080
OPROMOT	0.099	0.179	0.159	0.110
ISUP	0.154	0.177	0.140	0.280
PROSUP	0.213	0.212	0.038	0.260

**Sample Correlation Matrix (contd...)**

	ISUP	PROSUP
UNFAIR		
DCHARG		
UNEMP		
ITRAIN		
STRAIN		
ETRAIN		
IPROMOT		
OPROMOT		
ISUP	1.000	
PROSUP	0.475	1.000

Number of Cases : 213

**Reproduced Correlation Matrix**

	UNFAIR	DCHARG	UNEMP	ITRAIN
UNFAIR	1.000			
DCHARG	0.481	1.000		
UNEMP	0.382	0.602	1.000	
ITRAIN	0.081	0.128	0.102	1.000
STRAIN	0.093	0.146	0.116	0.638
ETRAIN	0.089	0.140	0.111	0.609
IPROMOT	0.140	0.221	0.176	0.171
OPROMOT	0.121	0.192	0.152	0.148
ISUP	0.124	0.196	0.156	0.364
PROSUP	0.098	0.154	0.122	0.286

**Reproduced Correlation Matrix (contd...)**

	ISUP	PROSUP
UNFAIR		
DCHARG		
UNEMP		
ITRAIN		
STRAIN		
ETRAIN		
IPROMOT		
OPROMOT		
ISUP	1.000	
PROSUP	0.560	1.000

lations : 1.138E-08

	STRAIN	ETRAIN	IPROMOT	OPROMOT
00				
13	1.000			
14	0.694	1.000		
32	0.240	0.237	1.000	
15	0.184	0.208	0.683	1.000
34	0.456	0.348	0.389	0.319
53	0.337	0.262	0.263	0.185

	STRAIN	ETRAIN	IPROMOT	OPROMOT
0				
3	1.000			
9	0.695	1.000		
1	0.195	0.186	1.000	
8	0.169	0.161	0.743	1.000
4	0.415	0.396	0.377	0.327
5	0.326	0.311	0.296	0.257



	UNFAIR	DCHARG	UNEMP
UNFAIR	0.000		
DCHARG	-0.043	0.000	
UNEMP	-0.133	-0.148	0.000
ITRAIN	0.068	-0.018	-0.045
STRAIN	0.080	0.062	-0.088
ETRAIN	0.095	0.028	-0.117
IPROMOT	-0.007	-0.011	-0.007
OPROMOT	-0.023	-0.013	0.007
ISUP	0.030	-0.020	-0.016
PROSUP	0.115	0.057	-0.084

### Residual Matrix (correlations) (co

	ISUP	PROSUP
UNFAIR		
DCHARG		
UNEMP		
ITRAIN		
STRAIN		
ETRAIN		
IPROMOT		
OPROMOT		
ISUP	0.000	
PROSUP	-0.085	0.000

Value of the Maximum Absolute Residual

### ADFU Estimates of Free Parameters :

Path	Parameter Number	P
UNFAIR <- JOBSEC	1	
DCHARG <- JOBSEC	2	
UNEMP <- JOBSEC	3	
ITRAIN <- TRAINING	4	
STRAIN <- TRAINING	5	
ETRAIN <- TRAINING	6	
IPROMOT <- PROMOT	7	
OPROMOT <- PROMOT	8	
ISUP <- RELSUP	9	
PROSUP <- RELSUP	10	
JOBSEC <- TRAINING	11	
JOBSEC <- PROMOT	12	
JOBSEC <- RELSUP	13	

### ADFU Estimates of Free Parameters

Path	Standard Error
UNFAIR <- JOBSEC	0.061
DCHARG <- JOBSEC	0.061
UNEMP <- JOBSEC	0.061
ITRAIN <- TRAINING	0.047
STRAIN <- TRAINING	0.028
ETRAIN <- TRAINING	0.035
IPROMOT <- PROMOT	0.052
OPROMOT <- PROMOT	0.054
	0.05

ITRAIN	STRAIN	ETRAIN	IFRMO1	CFRMO1
0.000				
-0.095	0.000			
-0.065	0.000	0.000		
-0.089	0.045	0.051	0.000	
-0.033	0.016	0.047	-0.060	0.000
-0.080	0.042	-0.047	0.011	-0.008
-0.023	0.012	-0.049	-0.034	-0.072

ntd...)

al : 0.148

### in Dependence Relationships

Point Estimate	90.00% Confidence Interval	
	Lower	Upper
0.552	0.451	0.653
0.871	0.770	0.972
0.692	0.592	0.791
0.748	0.670	0.826
0.853	0.808	0.899
0.814	0.756	0.873
0.926	0.842	1.011
0.802	0.714	0.891
0.844	0.752	0.937
0.663	0.568	0.758
0.074	-0.129	0.277
0.192	0.075	0.310
0.132	-0.081	0.345

### in Dependence Relationships (contd...)

t
9.011
14.218
11.416
15.780
30.814
23.035
17.981
14.965
14.968

Variable	Estimate
UNFAIR	1.008
DCHARG	0.962
UNEMP	0.974
ITRAIN	1.000
STRAIN	1.002
ETRAIN	0.983
IPROMOT	0.989
OPROMOT	1.001
ISUP	0.998
PROSUP	0.970

### Values of Fixed Parameters in Dependence

Path	Value
UNFAIR <- E1	1.000
DCHARG <- E2	1.000
UNEMP <- E3	1.000
ITRAIN <- E4	1.000
STRAIN <- E5	1.000
ETRAIN <- E6	1.000
IPROMOT <- E7	1.000
OPROMOT <- E8	1.000
ISUP <- E9	1.000
PROSUP <- E10	1.000
JOBSEC <- Z1	1.000

### ADFU Estimates of Free Parameters in Vari

Path	Parameter Number	Point Esti
TRAINING <-> PROMOT	14	0
TRAINING <-> RELSUP	15	0
PROMOT <-> RELSUP	16	0
E1 <-> E1	17	0
E2 <-> E2	18	0
E3 <-> E3	19	0
E4 <-> E4	20	0
E5 <-> E5	21	0
E6 <-> E6	22	0
E7 <-> E7	23	0
E8 <-> E8	24	0
E9 <-> E9	25	0
E10 <-> E10	26	0
Z1 <-> Z1	27	0

### ADFU Estimates of Free Parameters in Vari

Path	Standard Error
TRAINING <-> PROMOT	0.075 3.29
TRAINING <-> RELSUP	0.069 8.39
PROMOT <-> RELSUP	0.073 6.61
E1 <-> E1	0.068 10.28
E2 <-> E2	0.107 2.26
E3 <-> E3	0.084 6.22
E4 <-> E4	0.071 6.20
E5 <-> E5	0.047 5.75
E6 <-> E6	0.058 5.84
E7 <-> E7	0.095 1.48

## Relationships

### ance/Covariance Relationships

Estimate	90.00% Confidence Interval	
	Lower	Upper
0.246	0.120	0.364
0.576	0.452	0.677
0.482	0.354	0.593
0.695	0.593	0.816
0.242	0.117	0.499
0.522	0.400	0.679
0.440	0.338	0.574
0.272	0.204	0.362
0.337	0.254	0.446
0.142	0.047	0.429
0.356	0.239	0.530
0.287	0.166	0.495
0.560	0.448	0.702
0.898	0.818	0.945

### ance/Covariance Relationships (contd...)

t

7  
9  
8  
2  
4  
4  
6  
3  
9  
3



Path	Value
TRAINING <-> TRAINING	1.000
PROMOT <-> PROMOT	1.000
RELSUP <-> RELSUP	1.000

### Equality Constraints on Variances

Constraint	Value	Lag Multiplier
JOBSEC <-> JOBSEC	1.000	
UNFAIR <-> UNFAIR	1.000	
DCHARG <-> DCHARG	1.000	
UNEMP <-> UNEMP	1.000	
ITRAIN <-> ITRAIN	1.000	
STRAIN <-> STRAIN	1.000	
ETRAIN <-> ETRAIN	1.000	
IPROMOT <-> IPROMOT	1.000	
OPROMOT <-> OPROMOT	1.000	
ISUP <-> ISUP	1.000	
PROSUP <-> PROSUP	1.000	

### ADFU Discrepancy Function

#### Measures of Fit of the Model

Sample Discrepancy Function Value

Population Discrepancy Function Value

Bias Adjusted Point Estimate  
90% Confidence Interval

Root Mean Square Error of Approximation

Steiger-Lind :  $RMSEA = \sqrt{Fo/df}$   
Point Estimate (modified AIC)  
90% Confidence Interval

Expected Cross-Validation Index (CVI)

Point Estimate (modified AIC)  
90% Confidence Interval  
CVI (modified AIC) for the Saturated

Test Statistic

Exceedance Probabilities  
Ho: Perfect Fit ( $RMSEA = 0.0$ )  
Ho: Close Fit ( $RMSEA \leq 0.050$ )

Multiplier for Obtaining Test Statistic  
Degrees of Freedom  
Effective Number of Parameters

If the usual SYSTAT cases-by-variables are printed before the iteration details of normality assumptions. They can be used to check the appropriateness of the statistics and standard errors if the underlying distribution is appropriate.

range plier	Standard Error
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000
0.000	0.000

: 0.185 (0.185)

, Fo

: 0.048  
: (0.000, 0.144)

on (RMSEA)

: 0.041  
: (0.000, 0.071)

Model : 0.430  
: (0.382, 0.526)  
: 0.519

: 39.136

: 0.099  
: 0.662

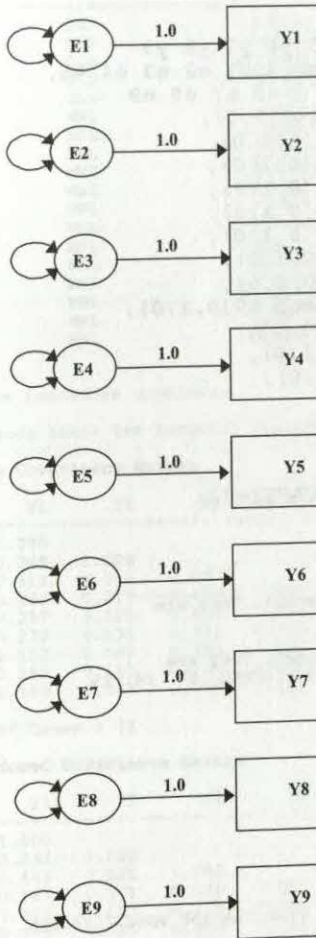
tic : 212.000  
: 29  
: 26

bles file is used as input, then the kurtosis estimates  
ls. These can be used to judge the appropriateness  
also be used to manually apply corrections to test  
ser is willing to accept that the assumption of an  
for the data (Shapiro and Browne, 1987).

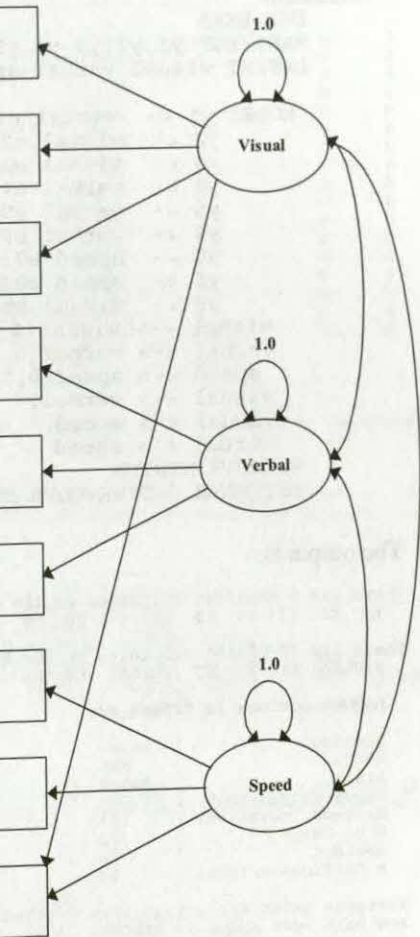
Lawley and Maxwell (1971) gave correct parameter estimates in a restricted factor analysis. This example shows how RAMONA can produce standard errors used for calculating the standard errors of the parameter estimates. RAMONA makes use of constrained optimization techniques in their formula by applying the delta method. It is shown, however, that the two methods are equivalent. In this example, Lawley and Maxwell made use of a sample correlation matrix administered to 72 children.

standard errors for maximum likelihood analysis model for a correlation matrix. This corrects these standard errors. The method differs from that of Lawley and Maxwell in that Lawley and Maxwell obtained unstandardized estimates. It can be shown, however, that the two methods will produce the same results. Lawley and Maxwell obtained a correlation matrix between nine ability tests





We analyze the relationships in the difference between the two runs is covariance structure and then as a c the first run and CORR in the second



path diagram using the correlation matrix. The that we first treat the model (inappropriately) as a correlation structure. We specify TYPE as COVA in ad.

RAMONA

USE EX4A

MANIFEST y1 y2 y3 y4 y5 y6

LATENT visual verbal speed

MODEL y1 <- visual e1(0,

y2 <- visual e2(0,

y3 <- visual e3(0,

y4 <- verbal e4(0,

y5 <- verbal e5(0,

y6 <- verbal e6(0,

y7 <- speed e7(0,1,

y8 <- speed e8(0,1,

y9 <- visual speed

visual <-> visual(0,1.0)

verbal <-> verbal(0,1.0)

speed <-> speed(0,1.0),

visual <-> verbal,

visual <-> speed,

verbal <-> speed

PLENGTH MEDIUM

ESTIMATE / TYPE=COVA NCASE

The output is:

There are 9 Manifest Variables in the Model.  
Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9

There are 12 Latent Variables in the Model.  
VISUAL E1 E2 E3 VERBAL E4 E5 E6 S

**RAMONA Options in Effect are**

Display	Covar
Method	MWL
Start	Rough
Convergence Limit	0.0001
Maximum Iterations	100
N of Cases	72
Restart	No
% Confidence Level	90

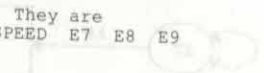
Variance paths for errors were omitted from  
and have been added by RAMONA.

Number of Manifest Variables : 9  
Total Number of Variables in the System : 2

Reading Correlation Matrix...

\*\*\* WARNING \*\*\* : A correlation matrix was  
standard errors may be inappropriate.

y7 y8 y9  
 e1 e2 e3 e4 e5,  
 6 e7 e8 e9  
 1.0),  
 1.0),  
 1.0),  
 1.0),  
 1.0),  
 1.0),  
 0),  
 0),  
 e9(0,1.0),  
 ),  
 ,



provided although DISP=COV fit measures and



Iteration	Method	Discr. Funct.
	OLS	1.013
1(0)	OLS	0.437
2(0)	OLS	0.144
3(0)	OLS	0.135
4(0)	OLS	0.135
4(0)	MWL	0.472
5(0)	MWL	0.426
6(0)	MWL	0.422
7(0)	MWL	0.421
8(0)	MWL	0.421
9(0)	MWL	0.421
10(0)	MWL	0.421
11(0)	MWL	0.421
12(0)	MWL	0.421
13(0)	MWL	0.421
14(0)	MWL	0.421
15(0)	MWL	0.421
16(0)	MWL	0.421

Iterative procedure complete.

Convergence Limit for Residual Cosine

### Sample Covariance Matrix

	Y1	Y2	Y3	Y4
Y1	1.000			
Y2	0.245	1.000		
Y3	0.418	0.362	1.000	
Y4	0.282	0.217	0.425	1.000
Y5	0.257	0.125	0.304	0.784
Y6	0.239	0.131	0.330	0.743
Y7	0.122	0.149	0.265	0.185
Y8	0.253	0.183	0.329	0.021
Y9	0.583	0.147	0.455	0.381

Number of Cases : 72

### Reproduced Covariance Matrix

	Y1	Y2	Y3	Y4
Y1	1.000			
Y2	0.232	1.000		
Y3	0.448	0.225	1.000	
Y4	0.341	0.171	0.330	1.000
Y5	0.325	0.163	0.315	0.788
Y6	0.309	0.155	0.300	0.748
Y7	0.210	0.105	0.203	0.052
Y8	0.298	0.149	0.289	0.074
Y9	0.517	0.260	0.501	0.351

### Residual Matrix (covariances)

	Y1	Y2	Y3	Y4
Y1	0.000			
Y2	0.013	0.000		
Y3	-0.030	0.137	0.000	
Y4	-0.059	0.046	0.095	0.000

Max.R.Cos.      Max.Const.      NRP      NBD

0.650	0	0
0.092	0	0
0.054	0	0
0.005	0	0
0.165	0	0
0.031	0	0
0.020	0	0
0.006	0	0
0.006	0	0
0.001	0	0
0.002	0	0
0.000	0	0
0.000	0	0
0.000	0	0
0.000	0	0
0.000	0	0
0.000	0	0

es: 1.000E-04 on 2 Consecutive Iterations

Y5      Y6      Y7      Y8      Y9

1.000				
0.730	1.000			
0.221	0.118	1.000		
0.139	-0.027	0.601	1.000	
0.400	0.235	0.385	0.462	1.000

Y5      Y6      Y7      Y8      Y9

1.000				
0.715	1.000			
0.050	0.047	1.000		
0.070	0.067	0.601	1.000	
0.336	0.319	0.331	0.471	1.000

Y4      Y5      Y6      Y7      Y8      Y9

.000  
0.004      0.000

# ML Estimates of Free Parameters in Dependent

Path	Parameter Number	Point Estimate
Y1 <- VISUAL	1	0.63
Y2 <- VISUAL	2	0.34
Y3 <- VISUAL	3	0.65
Y4 <- VERBAL	4	0.90
Y5 <- VERBAL	5	0.86
Y6 <- VERBAL	6	0.86
Y7 <- SPEED	7	0.65
Y8 <- SPEED	8	0.92
Y9 <- VISUAL	9	0.67
Y9 <- SPEED	10	0.19

# ML Estimates of Free Parameters in Dependent

Path	Standard Error	t
Y1 <- VISUAL	0.119	5.700
Y2 <- VISUAL	0.130	2.630
Y3 <- VISUAL	0.120	5.499
Y4 <- VERBAL	0.095	9.514
Y5 <- VERBAL	0.098	8.870
Y6 <- VERBAL	0.100	8.229
Y7 <- SPEED	0.131	4.973
Y8 <- SPEED	0.142	6.506
Y9 <- VISUAL	0.135	4.978
Y9 <- SPEED	0.130	1.471

# Values of Fixed Parameters in Dependence

Path	Value
Y1 <- E1	1.000
Y2 <- E2	1.000
Y3 <- E3	1.000
Y4 <- E4	1.000
Y5 <- E5	1.000
Y6 <- E6	1.000
Y7 <- E7	1.000
Y8 <- E8	1.000
Y9 <- E9	1.000

# ML Estimates of Free Parameters in Variance

Path	Parameter Number	Point Estimate
VISUAL <-> VERBAL	11	
VISUAL <-> SPEED	12	
VERBAL <-> SPEED	13	
E1 <-> E1	14	
E2 <-> E2	15	
E3 <-> E3	16	
E4 <-> E4	17	
E5 <-> E5	18	
E6 <-> E6	19	
E7 <-> E7	20	
E8 <-> E8	21	
E9 <-> E9	22	

# ndence Relationships

te	90.00% Confidence Interval	
	Lower	Upper
79	0.483	0.876
41	0.128	0.554
59	0.462	0.856
08	0.751	1.065
67	0.707	1.028
24	0.659	0.989
51	0.435	0.866
24	0.691	1.158
70	0.449	0.892
92	-0.023	0.406

# ndence Relationships (contd...)

# Relationships

# nance/Covariance Relationships

timate	90.00% Confidence Interval	
	Lower	Upper
0.552	0.344	0.708
0.474	0.210	0.674
0.088	-0.132	0.299
0.538	0.373	0.777
0.884	0.664	1.177
0.566	0.398	0.806
0.175	0.100	0.308
0.248	0.162	0.378
0.321	0.224	0.459
0.577	0.387	0.859
0.146	0.014	1.473
0.392	0.255	0.604



### ML Estimates of Free Parameters in

Path	Standard Error
VISUAL <-> VERBAL	0.111
VISUAL <-> SPEED	0.143
VERBAL <-> SPEED	0.133
E1 <-> E1	0.120
E2 <-> E2	0.154
E3 <-> E3	0.122
E4 <-> E4	0.060
E5 <-> E5	0.064
E6 <-> E6	0.070
E7 <-> E7	0.140
E8 <-> E8	0.205
E9 <-> E9	0.103

### Values of Fixed Parameters in Vari

Path	Value
VISUAL <-> VISUAL	1.000
VERBAL <-> VERBAL	1.000
SPEED <-> SPEED	1.000

### Maximum Likelihood Discrepancy Func

#### Measures of Fit of the Model

Sample Discrepancy Function Value

#### Population Discrepancy Function Val

Bias Adjusted Point Estimate  
90% Confidence Interval

#### Root Mean Square Error of Approximat

Steiger-Lind : RMSEA =  $\sqrt{Fo/df}$   
Point Estimate (modified AIC)  
90% Confidence Interval

#### Expected Cross-Validation Index (CV

Point Estimate (modified AIC)  
90% Confidence Interval  
CVI (modified AIC) for the Saturated

Test Statistic

Exceedance Probabilities  
Ho: Perfect Fit (RMSEA = 0.0)  
Ho: Close Fit (RMSEA  $\leq$  0.050)

Multiplier for Obtaining Test Statist  
Degrees of Freedom  
Effective Number of Parameters

## Analyzing the Correlation Structure

The maximum likelihood estimates  
the standard errors differ. Those fro  
errors in Lawley and Maxwell; tho

## Path Analysis (RAMONA)

## Variance/Covariance Relationships (contd...)

t

1 4.974  
 3 3.324  
 3 0.661  
 0 4.487  
 4 5.746  
 2 4.654  
 0 2.919  
 4 3.885  
 0 4.592  
 0 4.125  
 5 0.711  
 3 3.813

## Variance/Covariance Relationships

tion

: 0.421 (0.421)

ue, Fo

: 0.097  
 : (0.000, 0.354)

tion (RMSEA)

: 0.065  
 : (0.000, 0.124)

I)

: 1.041  
 : (0.944, 1.298)

d Model

: 1.268  
 : 29.891

: 0.153  
 : 0.330

stic

: 71.000  
 : 23  
 : 22

and measures of  $t$  from the two jobs are the same;  
 from the first job agree with the incorrect standard  
 from the second job agree with Lawley and

Maxwell's correct standard errors. A c shows that the introduction of additional multipliers (TYPE = CORR) results in a run differs from the first only in that w COVA.

The input is:

```

RAMONA
USE EX4B
MANIFEST y1 y2 y3 y4 y5
LATENT visual verbal speed

MODEL y1 <- visual e1
      y2 <- visual e2
      y3 <- visual e3
      y4 <- verbal e4
      y5 <- verbal e5
      y6 <- verbal e6
      y7 <- speed e7
      y8 <- speed e8
      y9 <- visual speed
visual <-> visual(0,1)
verbal <-> verbal(0,1)
speed <-> speed(0,1)
visual <-> verbal,
visual <-> speed,
verbal <-> speed
PLENGTH MEDIUM
ESTIMATE / TYPE=CORR

```

The output is:

There are 9 Manifest Variables in the Model  
Y1 Y2 Y3 Y4 Y5 Y6 Y7 Y8 Y9

There are 12 Latent Variables in the Model  
VISUAL E1 E2 E3 VERBAL E4 E5 E6 E7 E8 E9

#### RAMONA Options in Effect are

Display		Corr
Method		MWL
Start		Rough
Convergence Limit		0.0001
Maximum Iterations		100
N of Cases		72
Restart		No
% Confidence Level		90

Variance paths for errors were omitted and have been added by RAMONA.

Number of Manifest Variables  
Total Number of Variables in the System

Reading Correlation Matrix...

comparison of iteration times in the two jobs  
nal (nuisance) parameters and Lagrange  
substantially slower iteration times. The second  
ve specified TYPE = CORR instead of TYPE =

```
y6 y7 y8 y9  
eed e1 e2 e3 e4 e5,  
e6 e7 e8 e9  
(0,1.0),  
(0,1.0),  
(0,1.0),  
(0,1.0),  
(0,1.0),  
(0,1.0),  
(0,1.0),  
(0,1.0),  
(0,1.0),  
eed e9(0,1.0),  
1.0),  
.0),  
.0),
```

NCASES=72

Model. They are

Model. They are  
E6 SPEED E7 E8 E9

from the job specification

: 9  
n : 30



### Details of Iterations

Iteration	Method	Discr. Fun
0	OLS	1
1(0)	OLS	0
2(0)	OLS	0
3(0)	OLS	0
4(0)	OLS	0
4(0)	MWL	0
5(0)	MWL	0
6(0)	MWL	0
7(0)	MWL	0
8(0)	MWL	0
9(0)	MWL	0
10(0)	MWL	0
11(0)	MWL	0
12(0)	MWL	0
13(0)	MWL	0
14(0)	MWL	0
15(0)	MWL	0
16(0)	MWL	0

Iterative procedure complete.  
 Convergence Limit for Residual Co  
 Convergence Limit for Variance Co  
 Value of the Maximum Variance Co

### Sample Correlation Matrix

	Y1	Y2	Y3	
Y1	1.000			
Y2	0.245	1.000		
Y3	0.418	0.362	1.000	
Y4	0.282	0.217	0.425	1
Y5	0.257	0.125	0.304	0
Y6	0.239	0.131	0.330	0
Y7	0.122	0.149	0.265	0
Y8	0.253	0.183	0.329	0
Y9	0.583	0.147	0.455	0

Number of Cases : 72

### Reproduced Correlation Matrix

	Y1	Y2	Y3	
Y1	1.000			
Y2	0.232	1.000		
Y3	0.448	0.225	1.000	
Y4	0.341	0.171	0.330	1
Y5	0.325	0.163	0.315	0
Y6	0.309	0.155	0.300	0
Y7	0.210	0.105	0.203	0
Y8	0.298	0.149	0.289	0
Y9	0.517	0.260	0.501	0

### Residual Matrix (correlations)

	Y1	Y2	Y3	
Y1	0.000			
Y2	0.013	0.000		
Y3	-0.030	0.137	0.000	
Y4	-0.059	0.046	0.095	
Y5	-0.068	-0.038	-0.011	
Y6	-0.070	-0.024	0.030	
Y7	-0.088	0.044	0.062	
Y8	-0.045	0.034	0.040	
Y9	0.066	-0.113	-0.046	

## Path Analysis (RAMONA)

Fact.	Max.R.Cos.	Max.Const.	NRP	NBD
013		0.000		
437	0.650	0.193	0	0
144	0.092	0.018	0	0
135	0.054	0.007	0	0
135	0.005	0.000	0	0
472	0.165	0.000	0	0
426	0.031	0.003	0	0
422	0.020	0.001	0	0
421	0.006	0.000	0	0
421	0.006	0.000	0	0
421	0.001	0.000	0	0
421	0.002	0.000	0	0
421	0.000	0.000	0	0
421	0.000	0.000	0	0
421	0.000	0.000	0	0
421	0.000	0.000	0	0
421	0.000	0.000	0	0
421	0.000	0.000	0	0
421	0.000	0.000	0	0

osines: 1.000E-04 on 2 Consecutive Iterations  
 nstraint Violations: 5.000E-07  
 nstraint Violations : 2.142E-09

	Y4	Y5	Y6	Y7	Y8	Y9
000						
784	1.000					
743	0.730	1.000				
185	0.221	0.118	1.000			
021	0.139	-0.027	0.601	1.000		
381	0.400	0.235	0.385	0.462	1.000	

	Y4	Y5	Y6	Y7	Y8	Y9
000						
788	1.000					
748	0.715	1.000				
052	0.050	0.047	1.000			
074	0.070	0.067	0.601	1.000		
351	0.336	0.319	0.331	0.471	1.000	

	Y4	Y5	Y6	Y7	Y8	Y9
0.000						
-0.004	0.000					
-0.005	0.015	0.000				
0.133	0.171	0.071	0.000			
-0.053	0.069	-0.094	0.000	0.000		
0.030	0.064	-0.084	0.054	-0.009	0.000	

Value of the Maximum Absolute Residual

**ML Estimates of Free Parameters in**

Path	Parameter Number	Point Est
Y1 <- VISUAL	1	
Y2 <- VISUAL	2	
Y3 <- VISUAL	3	
Y4 <- VERBAL	4	
Y5 <- VERBAL	5	
Y6 <- VERBAL	6	
Y7 <- SPEED	7	
Y8 <- SPEED	8	
Y9 <- VISUAL	9	
Y9 <- SPEED	10	

**ML Estimates of Free Parameters in**

Path	Standard Error	
Y1 <- VISUAL	0.086	7.8
Y2 <- VISUAL	0.121	2.8
Y3 <- VISUAL	0.089	7.4
Y4 <- VERBAL	0.036	25.9
Y5 <- VERBAL	0.041	21.4
Y6 <- VERBAL	0.047	17.6
Y7 <- SPEED	0.103	6.2
Y8 <- SPEED	0.111	8.2
Y9 <- VISUAL	0.113	5.9
Y9 <- SPEED	0.129	1.4

**Scaled Standard Deviation (nuisance**

Variable	Estimate
Y1	1.000
Y2	1.000
Y3	1.000
Y4	1.000
Y5	1.000
Y6	1.000
Y7	1.000
Y8	1.000
Y9	1.000

**Values of Fixed Parameters in Depend**

Path	Value
Y1 <- E1	1.000
Y2 <- E2	1.000
Y3 <- E3	1.000
Y4 <- E4	1.000
Y5 <- E5	1.000
Y6 <- E6	1.000
Y7 <- E7	1.000
Y8 <- E8	1.000
Y9 <- E9	1.000

**ML Estimates of Free Parameters in V**

1 : 0.171

### Dependence Relationships

Estimate	90.00% Confidence Interval	
	Lower	Upper
0.679	0.537	0.822
0.341	0.143	0.539
0.659	0.513	0.804
0.908	0.850	0.967
0.867	0.801	0.934
0.824	0.747	0.901
0.651	0.480	0.821
0.924	0.741	1.108
0.670	0.485	0.856
0.192	-0.021	0.404

### Dependence Relationships (contd...)

t

868  
827  
442  
520  
414  
656  
291  
96  
957  
483

parameters)

### Dependence Relationships

### Variance/Covariance Relationships



Path	Parameter Number
VISUAL <-> VERBAL	11
VISUAL <-> SPEED	12
VERBAL <-> SPEED	13
E1 <-> E1	14
E2 <-> E2	15
E3 <-> E3	16
E4 <-> E4	17
E5 <-> E5	18
E6 <-> E6	19
E7 <-> E7	20
E8 <-> E8	21
E9 <-> E9	22

#### ML Estimates of Free Parameters

Path	Standard Error
VISUAL <-> VERBAL	0.
VISUAL <-> SPEED	0.
VERBAL <-> SPEED	0.
E1 <-> E1	0.
E2 <-> E2	0.
E3 <-> E3	0.
E4 <-> E4	0.
E5 <-> E5	0.
E6 <-> E6	0.
E7 <-> E7	0.
E8 <-> E8	0.
E9 <-> E9	0.

#### Values of Fixed Parameters in V

Path	Value
VISUAL <-> VISUAL	1.000
VERBAL <-> VERBAL	1.000
SPEED <-> SPEED	1.000

#### Equality Constraints on Variance

Constraint	Value	Lagrangian Multiplier
Y1 <-> Y1	1.000	0.00
Y2 <-> Y2	1.000	0.00
Y3 <-> Y3	1.000	0.00
Y4 <-> Y4	1.000	0.00
Y5 <-> Y5	1.000	0.00
Y6 <-> Y6	1.000	0.00
Y7 <-> Y7	1.000	0.00
Y8 <-> Y8	1.000	0.00
Y9 <-> Y9	1.000	0.00

#### Maximum Likelihood Discrepancy

#### Measures of Fit of the Model

Sample Discrepancy Function Value

Population Discrepancy Function

Bias Adjusted Point Estimate  
90% Confidence Interval

Root Mean Square Error of Approximation

## Path Analysis (RAMONA)

Point Estimate	90.00% Confidence Interval	
	Lower	Upper
0.552	0.344	0.708
0.474	0.210	0.674
0.088	-0.132	0.299
0.538	0.376	0.771
0.884	0.758	1.030
0.566	0.403	0.794
0.175	0.096	0.322
0.248	0.155	0.395
0.321	0.216	0.476
0.577	0.393	0.847
0.146	0.014	1.491
0.392	0.250	0.615

## in Variance/Covariance Relationships (contd...)

error	t
111	4.974
143	3.324
133	0.661
117	4.587
082	10.740
117	4.854
065	2.715
070	3.525
077	4.170
135	4.286
206	0.707
107	3.655

## Variance/Covariance Relationships

ces

Standard Error
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000
0.000

Function

Value, Fo : 0.421 (0.421)

Value, Fo

: 0.097  
: (0.000, 0.354)

ximation (RMSEA)

Steiger-Lind : RMSEA = SQRT(Fo/df)  
 Point Estimate (modified AIC)  
 90% Confidence Interval

**Expected Cross-Validation Index (CVI)**

Point Estimate (modified AIC)  
 90% Confidence Interval  
 CVI (modified AIC) for the Saturated Model

Test Statistic

Exceedance Probabilities  
 Ho: Perfect Fit (RMSEA = 0.0)  
 Ho: Close Fit (RMSEA ≤ 0.050)

Multiplier for Obtaining Test Statistic  
 Degrees of Freedom  
 Effective Number of Parameters

## Computation

### RAMONA's Model

Let  $\mathbf{v}_1$  be a  $p \times 1$  vector of manifest variables, and let

$$\mathbf{v} = \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \end{bmatrix}$$

be the  $t \times 1$  vector ( $t = p + m$ ) representing latent variables. Suppose that  $\mathbf{B}$  is a  $t \times t$  matrix corresponding to the directed arrows in the model. The element  $b_{ij}$ , will appear in the  $i$ th row and  $j$ th column of  $\mathbf{B}$ . The vector  $\mathbf{v}_2$  from  $\mathbf{v}$  by replacing all elements corresponding to latent variables. The vector  $\mathbf{v}_x$  consists of exogenous variables with no incoming paths in the system of directed paths represented by  $\mathbf{B}$ .

$$\mathbf{v} = \mathbf{B}\mathbf{v} + \mathbf{v}_x$$

The formulation of the model given in (11.1) is a special case of RAM (McArdle and McDonald, 1983). The first  $p$  elements of  $\mathbf{v}$ . Also, the non-null elements of  $\mathbf{B}$  are factors rather than residuals. Let

: 0.065  
: (0.000, 0.124)

: 1.041  
: (0.944, 1.298)  
Model : 1.268

: 29.891

: 0.153  
: 0.330

c : 71.000  
: 23  
: 22

variables,  $\mathbf{v}_2$  be an  $m \times 1$  vector of latent

(11-1)

representing all variables in the system, manifest and latent. The path coefficient from the  $j$ th element,  $\mathbf{v}_j$ , of  $\mathbf{v}$  to the  $i$ th element,  $\mathbf{v}_i$ , is the  $i$ th row,  $j$ th column of  $\mathbf{B}$ . Let  $\mathbf{v}_x$  be a  $t \times 1$  vector formed by replacing non-null rows of  $\mathbf{B}$  by zeros. Thus,  $\mathbf{v}_x$  is a vector of endogenous variables replaced by zeros. The path diagram is then given by:

(11-2)

Equation (11-1) differs only slightly from that of (11-2) (Jöreskog, 1984). All non-null elements of  $\mathbf{v}_x$  are also elements of  $\mathbf{v}_x$  can, in some situations, be common



$$\Phi = \text{Cov}(\mathbf{v}_x, \mathbf{v}_x')$$

be the  $t \times t$  covariance matrix of  $\mathbf{v}_x$  associated with two-headed arrows.  $\Phi$  will be associated with endogenous variables.

Let  $\Upsilon = \text{Cov}(\mathbf{v}, \mathbf{v}')$ . It follows that

$$\Upsilon = (\mathbf{I} - \mathbf{B})^{-1} \Phi (\mathbf{I} - \mathbf{B}')^{-1}$$

The manifest variable covariance matrix is  $\Upsilon$  (see equation (11-1)). Specifications of covariances by applying constraints to  $\Upsilon$ .

The structural model employed in RAMONA and  $\Phi$  are large matrices with many elements alone are stored in RAMONA. The computation of  $(\mathbf{I} - \mathbf{B})^{-1}$  and  $\Upsilon$  is done by RAMONA.

The covariance structure in equation (11-1) and Weeks (1980) in that there is no distinction between two.

Structural equation models are used in many published studies where the RAMONA fits a correlation structure  $\mathbf{v}_i^*$ , with unit variance to correspond to the latent variable taking

$$\mathbf{v}_i = \sigma_i \mathbf{v}_i^* \text{ for } i \leq p$$

where  $\sigma_i$  stands for the standard deviation of  $\mathbf{v}_i$  the same way as latent variables.  $\sigma_i$  is endogenous and fixed at unity if  $\mathbf{v}_i$  is treated in the same way as a path coefficient expressing the manifest variable

$$\Sigma = \mathbf{D}_\sigma \mathbf{P} \mathbf{D}_\sigma$$

where  $\mathbf{D}_\sigma$  is a diagonal matrix with the standardized manifest variable correlation matrix and standardized duplicate variables

of  $\mathbf{v}_x$ . Thus, the nonzero elements of  $\Phi$  are parameters shown in the path diagram. Null rows and columns of  $\Phi$  correspond to exogenous variables in  $\mathbf{v}$ .

from equation (11-2) that (McArdle and McDonald)

$$(11-3)$$

The matrix  $\Sigma = \text{Cov}(\mathbf{v}_1, \mathbf{v}_1')$  is the first  $p \times p$  submatrix of  $\Gamma$ . Specified values may be assigned to exogenous variable variances and covariances to appropriate diagonal elements of  $\Sigma$ .

The matrix  $\mathbf{B}$  defined by RAMONA is given in equation (11-3). Both  $\mathbf{B}$  and  $\Sigma$  have most of their elements equal to 0. Their nonzero elements are specified in RAMONA. Sparse matrix methods are used in the estimation of  $\Gamma$ . Details can be found in Mels (1989).

Equation (11-3) differs from a formulation of Bentler (1980) in that it is a single matrix,  $\mathbf{B}$ , for path coefficients instead of

several matrices. It is often fitted to sample correlation matrices. There are several ways in which this has been done incorrectly (Cudeck, 1989).

One way to fit the model by introducing a duplicate standardized variable,  $\mathbf{v}_i^*$ , corresponding to each manifest variable  $\mathbf{v}_i$ ,  $i \leq p$ , and then

fitting the model to the standard deviation of  $\mathbf{v}_i$ . The duplicate variables are treated in the same way as the original—with variances constrained to unity if they are exogenous and with covariances specified if they are exogenous. Also, the standard deviation,  $\sigma_i$ , of  $\mathbf{v}_i$  is specified as a path coefficient. This procedure is equivalent to fitting the model to the covariance matrix in the form

with the  $\sigma_i$ ,  $i \leq p$ , as diagonal elements, and  $\mathbf{P}$  is the matrix of path coefficients, which is treated as the covariance matrix of the  $\mathbf{v}_i^*$ ,  $i \leq p$ . Fitting the model to a sample correlation

matrix instead of a sample covariance matrix results in the estimates  $\hat{\sigma}_i$  being replaced by  $\hat{\sigma}_i s_i$ , where  $s_i$  is a sample standard deviation. These quantities are referred to as *Scaled Standard Deviations (nuisance parameters)* in the output. Other parameter estimates are not affected.

This approach involves the introduction of  $p$  additional parameters,  $\sigma_i$ , and  $p$  additional constraints on the variances of  $\nu_i^*$ . The number of degrees of freedom is not affected (unless some parameters or constraints are redundant), but computation time is increased because of the additional parameters and additional constraints.

## Algorithms

Let  $\gamma$  be the parameter vector and  $\Sigma = \Sigma(\gamma)$  the covariance structure. Parameter estimates are obtained by minimizing a discrepancy function,  $F(S, \Sigma(\gamma))$ , specified using METHOD. Alternatives are:

MWL Maximum Wishart likelihood.

$$F(S, \Sigma) = \ln|\Sigma| - \ln|S| + \text{tr}[S \Sigma^{-1}] - p$$

GLS Generalized least squares assuming a Wishart distribution for  $S$ .

$$F(S, \Sigma) = \frac{1}{2} \text{tr}[S^{-1}(S - \Sigma)]^2$$

OLS Ordinary least squares.

$$F(S, \Sigma) = \frac{1}{2} \text{tr}[(S - \Sigma)]^2$$

ADFU, ADFG Asymptotically distribution-free methods

$$F(S, \Sigma) = (s - \sigma)' \Gamma^{-1} (s - \sigma)$$

where  $s$  and  $\sigma$  are column vectors with  $p(p+1)/2$  elements formed from the distinct elements of  $S$  and  $\Sigma$ , respectively, and  $\Gamma$  is an estimate of the asymptotic covariance matrix of sample covariances. For ADFU,  $\Gamma$  is unbiased (Browne, 1982) but need not be positive definite. If  $\Gamma$  is indefinite, the program moves automatically from ADFU to ADFG. With ADFG,  $\Gamma$  is biased but Gramian (Browne, 1982).

An iterative Gauss-Newton computing procedure with constraints (Browne and Du Toit, 1992) is used to obtain parameter estimates. With MWL, the weight matrix is re-

specified on each iteration. The procedure is then equivalent to the Aitchison and Silvey (1960) adaptation of the Fisher scoring method to deal with equality constraints.

Some computer programs can yield negative estimates of variances. This does not happen with RAMONA. Bounds are imposed to ensure that variance estimates are non-negative and that all correlation estimates lie between  $-1$  and  $+1$ . The imposition of these bounds can result in the convergence of RAMONA in situations where programs that do not impose them fail to converge. In some cases, a program that allows negative variance estimates and does converge will yield a smaller discrepancy function value than RAMONA.

Iteration is continued until the largest absolute residual cosine (Browne, 1982) falls below a tolerance, specified in CONV, on two consecutive iterations.

### **Confidence Intervals**

Approximate 90% confidence intervals are given for parameter estimates associated with dependence paths and with covariance paths. Confidence intervals for path coefficients and covariances (variances unrestricted) are provided under the assumption of a normal distribution for the estimator  $\hat{\gamma}$  (Browne, 1974) and are symmetric about the parameter estimate. Confidence intervals for other parameters are nonsymmetric about the parameter estimate (Browne, 1974) and are obtained under the following assumptions:

- Correlation coefficients (covariances with both corresponding variances restricted to unity): a normal distribution is assumed for the  $z$ -transform,  $\frac{1}{2} \ln[(1 + \hat{\gamma})/(1 - \hat{\gamma})]$ , (Browne, 1974).
- Variances: a normal distribution is assumed for the natural logarithm,  $\ln \hat{\gamma}$ , (Browne, 1974).
- Error variances under a correlation structure (corresponding dependent variable variances are constrained to unity): a normal distribution is assumed for  $-\ln(\hat{\gamma}^{-1} - 1)$  (Browne, 1974).

### **Measures of Fit of a Model**

This section provides a brief description of the measures of fit output by RAMONA. Further information concerning these measures of fit can be found in Browne and Cudeck (1993).



Let  $N = n + 1$  be the sample size;  $p$ , the number of manifest variables; and  $q$ , the number of free parameters in the model. Then the number of degrees of freedom is  $d = \frac{1}{2} p(p + 1) - q$ . The sample covariance matrix is denoted by  $S$  and the corresponding population covariance matrix by  $\Sigma_0$ .

The minimal sample discrepancy function value is:

$$\hat{F} = \text{Min}_{\gamma} F(S, \Sigma(\gamma))$$

and the corresponding minimal population discrepancy function value is:

$$F = \text{Min}_{\gamma} F(\Sigma_0, \Sigma(\gamma))$$

Now  $F_0$  is bounded below by 0 and takes on a value of 0 if and only if  $\Sigma_0$  satisfies the structural model exactly. Therefore, we can regard  $F_0$  as a measure of badness-of-fit of the model,  $\Sigma(\gamma)$ , to the population covariance matrix,  $\Sigma_0$ .

We assume that the test statistic  $n \hat{F}$  has an approximate noncentral chi-square distribution with  $d$  degrees of freedom and a noncentrality parameter  $\sigma = nF_0$ . This will be true if the discrepancy function is correctly specified for the distribution of the data,  $F_0$  is small enough, and  $N$  is large enough (Steiger, Shapiro, and Browne, 1985). Then the expected value of  $\hat{F}$  will be approximately  $F_0 + d/n$ , so that  $\hat{F}$  is a biased estimator of  $F_0$ . As a less biased point estimator of  $F_0$  we use:

$$\hat{F}_0 = \text{Max} \{ \hat{F} - (d/n), 0 \}$$

We also provide a 90% confidence interval on  $F_0$  as suggested by Steiger and Lind (1980). Let  $\Phi(x | \delta, d)$  be the cumulative distribution function of a noncentral chi-square distribution with noncentrality parameter  $\delta$  and  $d$  degrees of freedom. Given  $x = n \times \hat{F}$  and  $d$ , the lower limit,  $\delta_L$ , of the 90% confidence interval on  $n \times F_0$  is the solution for  $\delta$  of the equation

$$\Phi(x | \delta, d) = 0.95$$

and the upper limit  $\delta_U$  is the solution for  $\delta$  of

$$\Phi(x | \delta, d) = 0.05$$

A 90% confidence interval on  $F_0$  is then given by  $(n^{-1}\delta_L; n^{-1}\delta_U)$ .

Because  $F_0$  cannot increase if additional parameters are added, it gives little guidance about when to stop adding parameters. It is preferable to use the root mean square error of approximation (Steiger and Lind, 1980):

$$\text{RMSEA} = \sqrt{\frac{\hat{F}_0}{d}}$$

as a measure of the fit per degree of freedom of the model. This population measure of badness-of-fit is also bounded below by 0 and will be 0 only if the model fits perfectly. It will decrease if the inclusion of additional parameters substantially reduces  $F_0$  but will increase if the inclusion of additional parameters reduces  $F_0$  only slightly. Consequently, it can give some guidance as to how many parameters to use. Practical experience has suggested that a value of the RMSEA of about 0.05 or less indicates a close fit of the model in relation to the degrees of freedom. A value of about 0.08 or less indicates a reasonable fit of the model in relation to the degrees of freedom.

A point estimate of the RMSEA is given by:

$$\text{Estimate (RMSEA)} = \sqrt{\frac{\hat{F}_0}{d}}$$

and a 90% confidence interval by:

$$\text{Interval Estimate (RMSEA)} = \left( \sqrt{\frac{\delta_L}{nd}}, \sqrt{\frac{\delta_U}{nd}} \right) \quad (11-4)$$

The RMSEA does not depend on sample size and therefore does not take into account the fact that it is unwise to fit a model with many parameters if  $N$  is small. A measure of fit that does this is the expected cross-validation index (ECVI). Consider two samples of size  $N$ —a calibration sample  $C$  and a validation sample  $V$ . Suppose that the model is fitted to the calibration sample yielding a reproduced covariance matrix  $\hat{\Sigma}_C$ . The discrepancy between  $\hat{\Sigma}_C$  and the validation sample covariance matrix  $S_V$  is then measured with the discrepancy function yielding  $F(S_V, \hat{\Sigma}_C)$  as a measure of stability under cross-validation. A difficulty with this approach is that two samples are required. One can avoid a second sample by estimating the expected value of  $F(S_V, \hat{\Sigma}_C)$  from a single sample. Assume that the discrepancy function is correctly specified for the

distribution of the data. Taking expectations over calibration samples and validation samples gives the expected cross-validation index:

$$ECVI = \xi \xi_{CV} F(SV, \hat{\Sigma}_C) \approx F_0 + (d + 2q)/n \quad (11-5)$$

A point estimate of the ECVI is given by (Browne and Cudeck, 1990):

$$\text{Estimate (ECVI)} = \hat{F} + 2q/n \quad (11-6)$$

If METHOD is set to MWL, this point estimate of the ECVI is related by a linear transformation to the Akaike Information Criterion (Akaike, 1973) and will lead to the same conclusions.

The point estimate in equation (11-6) will decrease if an additional parameter reduces  $\hat{F}$  sufficiently and increases otherwise. This will give some guidance as to the number of parameters to retain. However, the amount of reduction in  $\hat{F}$  required before an increase in the point estimate occurs is affected by the sample size. If  $n$  is very large, increasing the number of parameters will tend to reduce the point estimate of the ECVI. One should also bear in mind that sampling variability affects the point estimates.

An approximate 90% confidence interval on the ECVI may be obtained from:

$$\text{Interval Estimate (ECVI)} = \left( \frac{\delta_L + d + 2q}{n}, \frac{\delta_U + d + 2q}{n} \right) \quad (11-7)$$

It can happen that  $(\hat{F} - d) < \delta_L$ , so that the point estimate in equation (11-6) is smaller than the lower limit of the confidence interval in equation (11-7). In particular, this will be true if the (approximately unbiased) point estimate in equation (11-6) is less than the lower bound  $(d + 2q)/n$  for the approximation to the ECVI given in equation (11-5).

For comparative purposes, RAMONA also provides the ECVI of the saturated model where no structure is imposed on  $\Sigma$ :

$$ECVI (\text{Saturated Model}) = \frac{2 \times (d + q)}{n}$$



The test statistic  $n \times F'$  is also output by RAMONA. We follow convention in providing the exceedance probability,  $1 - \Phi(n\hat{F} | 0, d)$ , for a test of the point hypothesis

$$H_0: F_0 = 0 \quad (11-8)$$

which implies that the model holds exactly. Our opinion, however, is that this null hypothesis is implausible and that it does not much help to know whether or not the statistical test has been able to detect that it is false. More relevant is the exceedance probability for an interval hypothesis of close fit, which we define by

$$H_0: \text{RMSEA} \leq 0.05 \quad (11-9)$$

and which implies that  $\delta \leq \delta^* = n \times d \times 0.05^2$ .

The exceedance probability output by RAMONA is given by  $1 - \Phi(n\hat{F} | \delta^*, d)$ .

Note that the null hypothesis of perfect fit in equation (11-8) is not rejected at the 5% level if  $\delta_L = 0$  or, equivalently, the lower limit of the confidence interval in equation (11-4) is 0. The null hypothesis of a close fit in equation (11-9) is not rejected at the 5% level if the lower limit of the confidence interval in equation (11-4) is not greater than 0.05.

When METHOD is set to MWL, two sets of measures of fit are output. One is based on the maximum likelihood discrepancy function value

$$\hat{F} = \ln|\hat{\Sigma}| - \ln|S| + \text{tr}[S\hat{\Sigma}^{-1}] - p$$

and the other on the generalized least squares discrepancy function value

$$\hat{F} = \frac{1}{2} \text{tr}[\hat{\Sigma}^{-1}(S - \hat{\Sigma})]$$

When the model fits well, the differences between the two sets of fit measures should be small (Browne, 1974).



## References

- Aitchison, J. and Silvey, S. D. (1960). Maximum likelihood estimation procedures and associated tests of significance. *Journal of the Royal Statistical Society, Series B*, 22, 154–171.
- Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petrov and F. Csaki (eds), *Proceedings of the Second International Symposium on Information Theory*, 267–281. Budapest: Akademiai Kiado.
- Bentler, P. M. and Weeks, D. G. (1980). Linear structural equations with latent variables. *Psychometrika*, 45, 289–308.
- Bollen, K. A. (1989). *Structural equations with latent variables*. New York: John Wiley & Sons.
- Browne, M. W. (1974). Generalized least squares estimators in the analysis of covariance structures. *South African Statistical Journal*, 8, 1–24. (Reprinted in Aigner, D.J. and Goldberger, A.S. (eds), *Latent Variables in Socioeconomic Models*, 205–226. Amsterdam: North Holland.)
- Browne, M. W. (1982). Covariance structures. In D. M. Hawkins (ed), *Topics in Applied Multivariate Analysis*, 72–141. Cambridge: Cambridge University Press.
- Browne, M. W. and Cudeck, R. (1990). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, 24, 445–455.
- Browne, M. W. and Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen and J. S. Long (eds), *Testing Structural Equation Models*, Beverly Hills, Calif.: Sage.
- Browne, M. W. and Du Toit, S. H. C. (1992). Automated fitting of nonstandard models. *Multivariate Behavioral Research*, 27, 269–300.
- Cudeck, R. (1989). Analysis of correlation matrices using covariance structure models. *Psychological Bulletin*, 105, 317–327.
- Duncan, O. D., Haller, A. O., and Portes, A. (1971). Peer influence on aspirations, a reinterpretation. *Causal Models in the Social Sciences*, H. M. Blalock, ed. 219–244. Aldine-Atherstone.
- Everitt, B. S. (1984). *An introduction to latent variable models*. London: Chapman and Hall.
- \*Guttman, L. (1954). A new approach to factor analysis: The radex. In P. F. Lazarsfeld (ed) *Mathematical Thinking in the Social Sciences*, 258–348. Glencoe: The Free Press.
- Jöreskog, K. G. (1977). Structural equation models in the social sciences: Specification estimation and testing. In P. R. Krishnaiah (ed), *Applications of Statistics*, 265–287. Amsterdam: North Holland.

- Lawley, D. N. and Maxwell, A. E. (1971). *Factor analysis as a statistical method*. 2nd ed. New York: American Elsevier.
- McArdle, J. J. (1988). Dynamic but structural equation modeling of repeated measures data. In J. R. Nesselroade and R. B. Cattell (eds), *The Handbook of Multivariate Experimental Psychology*, 2nd ed., 561–614. New York: Plenum Press.
- McArdle, J. J., and McDonald, R. P. (1984). Some algebraic properties of the Reticular Action Model for moment structures. *British Journal of Mathematical and Statistical Psychology*, 37, 234–251.
- McDonald, R. P. (1985). *Factor analysis and related methods*. Hillsdale: Erlbaum.
- Mels, G. (1989). *A general system for path analysis with latent variables*. M. S. thesis, University of South Africa.
- Mels, G. and Koorts, A. S. (1989). *Causal Models for various job aspects*. SAIPA, 24, 144–156.
- Shapiro, A. and Browne, M. W. (1987). Analysis of covariance structures under elliptical distributions. *Journal of the American Statistical Association*, 82, 1092–1097.
- Steiger, J. H. and Lind, J. C. (1980). *Statistically based tests for the number of common factors*. Paper presented at the annual meeting of the Psychometric Society: Iowa City.
- Steiger, J. H., Shapiro, A., and Browne, M. W. (1985). On the asymptotic distribution of sequential chi-square statistics. *Psychometrika*, 50, 253–264.
- Wheaton, B., Muthén, B., Alwin, D. F., and Summers, G. F. (1977). Assessing reliability and stability in panel models. In D. R. Heise (ed), *Sociological Methodology*, 84–136. San Francisco: Jossey-Bass.

(\*indicates additional reference.)

## Acknowledgments

The development of this program was partially supported by the Institute of Statistical Research of the South African Human Sciences Research Council, the South African Foundation for Research Development, and the University of South Africa.

The authors are indebted to Professor S.H.C. Du Toit and to Mrs. Yvette Seymore for a number of subroutines used in the program.

# Acronym & Abbreviation Expansions

## A

ABS - absolute value  
ACF - autocorrelation function  
ACOLOR - color axes  
ACS - arccosine  
ACT - actuarial life table  
AD test - Anderson Darling test  
ADDTREE - additive trees  
ADFG - asymptotically distribution free estimate biased, Gramian  
ADFU - asymptotically distribution free estimate unbiased  
ADJSEASON - seasonal adjustment  
AHMAX - maximum extent  
AHMIN - minimum extent  
AIC - Akaike information criterion  
AID - automatic interaction detection  
ALT - alternative  
ANCOVA - analysis of covariance  
ANG1 - deviation of angles from north in a clockwise direction  
ANG2 - deviation of angles from horizontal (for 3D models)  
ANG3 - tilt angle  
ANOVA - analysis of variance  
ANOVAHYPO - hypothesis tests in analysis of variance  
AR - autoregressive  
ARIMA - autoregressive integrated moving average  
ARL - average run length

ARMA - autoregressive moving average  
ARS - adaptive rejection sampling  
ASCII - American Standard Code for Information Interchange  
ASE - asymptotic standard error  
ASN - arcsine  
ATH - arc hyperbolic tangent  
ATN - arctangent  
AVERT - vertical extent  
AVG - average

## B

BC - Bray-Curtis similarity measure  
BCa - Bias Corrected and accelerated  
BCF - Beta cumulative function  
BDF - Beta density function  
BETACORR - beta correction  
BIC - Bayesian information criterion  
BIF - Beta inverse function  
BMP - Windows bitmap  
BOF - beginning-of-file  
BOG - beginning-of-BY group  
BONF - Bonferroni  
BOOT - bootstrap  
BRN - Beta random number

## C

CART - classification and regression trees  
CBSTAT - column basic statistics  
CCF - Cauchy cumulative function  
CCF - cross-correlation function  
CDF - Cauchy density function  
cdf/CF - cumulative distribution function  
CDFUNC - coefficients for canonical variables



CFUNC - coefficients for the classification functions  
 CGM - Computer graphics metafile: binary or clear text  
 CHAZ - cumulative hazard  
 CHISQ - Chi-square distribution  
 CHOL - Cholesky decomposition  
 CI - confidence interval  
 CIF - Cauchy inverse function  
 CIM - confidence interval of mean  
 CLASS - classification  
 CLSTEM - stem and leaf plot for column  
 CMeans - canonical scores of group means  
 CMULTIVAR - multiple string variables  
 COEF - coefficients  
 COL/col - column  
 COLPCT - Column percentages  
 CONFIG - configuration  
 CONT - Contingency coefficient  
 CONV - convergence  
 CORAN - correspondence analysis  
 CORR - correlations  
 CORR1 - single correlation coefficient  
 CORR2 - equality of two correlations  
 COV - covariance  
 Cp - process capability index  
 CPL - process capability based on lower specification limit  
 CPU - process capability based on upper specification limit  
 Cpk-Process capability index for off-centered process  
 CR - confidence region  
 CRA - cost of response above UTL  
 CRB - cost of response below LTL  
 CRN - Cauchy random number  
 CSCORE - canonical scores  
 CSIZE - size of characters  
 CSQ - Chi-square  
 CSTATISTICS - column statistics  
 CSV - comma separated values

CUSUM - cumulative sum  
 CUSUM HI - Upper cumulative sum  
 CUSUM LO - Lower cumulative sum  
 CV - coefficient of variation  
 CVI - cross validation index

## D

DBF - Dbase files  
 DC - deciles of risk  
 DECF - Double exponential cumulative function  
 DEDF - Double exponential density function  
 DEIF - Double exponential inverse function  
 DENFUN - density function  
 dep. - dependent  
 DERN - Double exponential random number  
 DET - determinant  
 DEVI - deviates (observed values - expected values)  
 DEXP - Double exponential distribution  
 df - degrees of freedom  
 DF - distribution function  
 DHAT - estimated distance  
 DIF - data interchange format  
 DIM - dimension  
 DISCRIM - discriminant analysis  
 DIST - distance  
 DIT - dot histogram  
 DOE - design of experiments  
 DOS - disc operating system  
 DPMO - defects per million opportunities  
 DPU - defects per unit  
 DTA - Stata files  
 DUCF - Discrete uniform cumulative function  
 DUDF - Discrete uniform density function  
 DUIF - Discrete uniform inverse function  
 DUNIFORM - Discrete uniform  
 DURN - Discrete uniform random number  
 DWLS - distance weighted least-squares

## E

ECF - Exponential cumulative function



EDF - Exponential density function  
 EEXP - extreme value exponential  
 EIF - Exponential inverse function  
 EIGEN - eigenvalues  
 ELAMBDA -  $\exp(\lambda)$   
 EM - expectation-maximization  
 EMF - Windows enhanced metafile  
 ENCF - Logit normal cumulative function  
 ENDF - Logit normal density function  
 ENIF - Logit normal inverse function  
 ENORMAL - Logit normal  
 ENRN - Logit normal random number  
 EOF - end-of-file  
 EOG - end-of-BY group  
 EPS - Encapsulated postscript  
 ERN - Exponential random number  
 ES - exhaustive search  
 ESS - error sum of squares  
 EW - extreme value Weibull  
 EWMA - exponentially weighted moving average  
 EXP/exp - exponential/ expected

## F

FAR - false-alarm rates  
 FCF - F cumulative function  
 FCOLOR - color foreground  
 FDF - F density function  
 FIF - F inverse function  
 FINV - inverse of the F cumulative  
 FITC - fitting distribution: continuous  
 FITD - fitting distribution: discrete  
 FITDIST - fitting distributions  
 Flexibeta - flexible beta  
 FPLOT - function plots  
 FRN - F random number  
 FTD - folded trellis detector  
 FTDEV - Freeman-Tukey deviate  
 FULLCOND - full conditional  
 FUN - function

## G

GCF - Gamma cumulative function  
 GCOR - groupwise correlation matrix  
 GCOV - groupwise covariance matrix  
 GCV - generalized cross validation  
 GDF - Gamma density function  
 GECF - Geometric cumulative function  
 GEDF - Geometric density function  
 GEIF - Geometric inverse function  
 GEN - general Toeplitz structure  
 GERN - Geometric random number  
 GG - Greenhouse Geisser  
 GIF - Gamma inverse function  
 GIF - Graphics Interchange Format  
 GLM - generalized linear models  
 GLMHYPO - hypothesis tests in general linear model  
 GLMPOST - post hoc estimate for repeated measures in general linear model  
 GLS - generalized least-squares  
 GMA - geometric moving average  
 GN - Gauss-Newton method  
 GOCHF - Gompertz cumulative function  
 GODF - Gompertz density function  
 GOIF - Gompertz inverse function  
 GORN - Gompertz random number  
 GRN - Gamma random number  
 GUCF - Gumbell cumulative function  
 GUDF - Gumbell density function  
 GUIF - Gumbell inverse function  
 GURN - Gumbell random number

## H

H & L - Hosmer and Lemeshow  
 HC - heteroscedasticity-consistent  
 HCF - Hypergeometric cumulative function  
 HDF - Hypergeometric density function  
 HF - Huynh-Feldt  
 HGEOMETRIC - hypergeometric  
 HIF - Hypergeometric inverse function  
 HIST - histogram  
 HKB - Hoerl, Kennard, and Baldwin

H-L trace - Holding-Lawley trace

HR - hit-rates

HRN - Hypergeometric random number

HSD - honestly significant differences

HTERM - terms tested hierarchically

HTML - hyper text markup language

HYMH - hybrid Metropolis-Hastings

## I

IF - Inverse cumulative distribution function

IGAUSSIAN - inverse Gaussian

IGCF - Inverse Gaussian cumulative function

IGDF - Inverse Gaussian density function

IGIF - Inverse Gaussian inverse function

IGRN - Inverse Gaussian random number

IIDMC - independently and identically distributed Monte Carlo

IMPSAMPI - importance sampling integration

IMPSAMPR - importance sampling ratio

I-MR - individual and moving range

Ind/indep - independent

IndMH - Independent Metropolis-Hastings

INDSCAL - individual differences scaling

INITSAMP - initial sample

INTEG FUN - integrated function

IPA - iterated principal axis

ITER - iterations

## J

JACK - jackknife

JCLASS - jackknifed classification

JMP - JMP v3.2 data files

JPEG/JPG - joint photographic experts group

## K

K-M - Kaplan-Meier

KNBD - kth nearest neighborhood

KRON - Kronecker product

K-S test - Kolmogorov-Smirnov test

KS1 - one sample Kolmogorov-Smirnov tests

KS2 - two sample Kolmogorov-Smirnov tests

## L

LAD - least absolute deviations

LB - larger the better

LCF - Logistic cumulative function

LCHAZ - log cumulative hazard

LCL - lower control limit

LCONV - log-likelihood convergence criteria

LDF - Logistic density function

LGM - log gamma

LGST - logistic

LIF - Logistic inverse function

L-L/LL - log likelihood

LMS - least median of squares

LMSREG - least median of squares regression

LNCF - Lognormal cumulative function

LNDF - Lognormal density function

LNIF - Lognormal inverse function

LNOR/LNORMAL - lognormal

LNRN - Lognormal random number

loc - location

LOG1 - one-parameter logistic (Rasch)

LOG2 - two-parameter logistic

LOGIT - logistic regression

LOGITHYPO - hypothesis tests in logistic regression

LOGLIN - loglinear modeling

LR - likelihood ratio

LRCHI - likelihood ratio chi-square

LRDEV - likelihood ratio of deviate

LRN - Logistic random number

LS - least-squares

LSD - least significant difference

LSL - lower specification limit

LSQ - least-squares

LTAB - life tables

LTL - lower tolerance limit

LW - Lawless and Wang

## M

MA - moving average



- MAD - mean absolute deviation  
 MAHAL - Mahalanobis distances  
 MANCOVA - multivariate analysis of covariance  
 MANOVA - multivariate analysis of variance  
 MANOVAHYPO - hypothesis tests in MANOVA  
 MANOVAPOST - post hoc estimate for repeated measures in MANOVA  
 MAR - missing at random  
 MAX - maximum  
 MAXSTEP - maximum number of steps  
 MCAR - missing completely at random  
 MCMC - Markov Chain Monte Carlo  
 MDPREF - multidimensional preference  
 MDS - multidimensional scaling  
 MIN - minimum  
 M-H- Metropolis-Hastings  
 MIS - number of missing values  
 MIX - mixed regression  
 MIXHIER - mixed regression for data having a hierarchical structure  
 MIXMULTY - mixed regression for data having a multivariate structure  
 ML - Maximum Likelihood  
 MLA - maximum likelihood analysis  
 MLE - maximum likelihood estimate  
 MML - maximum marginal likelihood  
 MRC - Multiple Regression and Correlation  
 MS - mean squares  
 MSE - mean square error  
 MSIGMA - sigma measurement  
 MT - Mersenne-Twister  
 MTW - MINITAB v11 data files  
 MU2 - Guttman's mu2 monotonicity coefficients  
 MULTIVAR - multiple variables  
 MW - minimum within sum of squares deviations  
 MWL - maximum Wishart likelihood  
  
 N  
 NAR - non-stationary first-order autoregressive  
 NB - nominal the best  
 NBB - nominal-the-best: bilateral tolerance  
 NBCF - Negative binomial cumulative function  
 NBD - number of active bounds on parameter values  
 NBDF - Negative binomial density function  
 NBIF - Negative binomial inverse function  
 NBINOMIAL - Negative binomial  
 NBRN - Negative binomial random number  
 NBU - nominal-the-best: unilateral tolerance  
 NCAT - number of categories  
 NCF - Binomial cumulative function  
 NCOL - number of columns  
 NDF - Binomial density function  
 NDMAX - maximum number of points  
 NDMIN - minimum number of points  
 NEM - number of EM iterations  
 NEXPO - negative exponential  
 NIF - Binomial inverse function  
 NIPALS - Nonlinear iterative partial least Squares  
 NLAG - number of lags  
 NLLOSS - nonlinear loss functions  
 NLMODEL - nonlinear models  
 NMIN - minimum count  
 NMULTIVAR - multiple numeric variables  
 NONLIN - nonlinear models  
 NP-Number nonconforming  
 NPAR - nonparametric  
 NREC - non-recreationist  
 NRN - Binomial random number  
 NROW - number of rows  
 NRP - number of apparently redundant parameters  
 NSAMP - number of sub-samples  
 NSPLIT - maximum number of splits  
 NX - number of nodes along the x axis  
 NXDIS - number of discretization points in the x (North) direction  
 NY - number of nodes along the y axis  
 NYDIS - number of discretization points in the y (East) direction  
 NZ - number of nodes along the z axis

NZDIS - number of discretization points in the z (Depth) direction

## O

Obs-observed

OBSFREQ - observed frequency

OC - operating characteristic

ODBC - open database capture and connectivity

OFREQ - outlier frequencies

OLS - ordinary least-squares

ORTHEQ- Equally Spaced Orthogonal component

ORTHUN- Unequally Spaced Orthogonal component

## P

P - Proportion nonconforming

PACF - Pareto cumulative function

PACF - partial autocorrelation function

PADF - Pareto density function

PAIF - Pareto inverse function

PARAM - parameters

PARN - Pareto random number

PCA - process capability analysis

PCF - iterated principal axis factoring

PCF - Poisson cumulative function

PCNTCHANGE - percentage change

PCT - Macintosh PICT

PDF - Poisson density function

pdf - probability density function

PDL - polynomial distributed lag

PERMAP - perceptual mapping

PIF - Poisson inverse function

PLIMITS - probability limits

PLS - partial least squares

pmf - probability mass function

PMIN - minimum proportion

PNG - Portable Network Graphics

POLY - polygon

POSAC - partially ordered scalogram analysis with coordinates

P-P - probability plot

PP - process performance

Ppk - Process performance index for off-centered process

PPL - process performance based on lower specification limit

PPM - parts per million

PPU - process performance based on upper specification limit

PRE - percentage reduction error

PREFMAP - preference mapping

PRN - Poisson random number

PROB - probability

PROP1 - single proportion

PROP2 - equality of two proportions

PS - PostScript

PVAF/p.v.a.f. -- present value annuity factor

p-value - probability value

## Q

QC - quality control

QMLE - quasi maximum likelihood estimate

QNTL - quantiles

QPLOT - quantile plots

Q-QPLOT - two sample quantile plot

QRD - QR decomposition

QS - quick search

QSK - quantitative symmetric similarity coefficients (or Kulczynski measure)

QUASI - Quasi-Newton method

## R

R & R - repeatability and reproducibility

R chart - range chart

RADMAX - maximum horizontal direction for the search radius

RADMIN - minimum horizontal direction for the search radius

RAND - random

RANDSAMP - random sampling

RANKREG - rank regression



RBSTAT - row basic statistics  
 RCF - Rayleigh cumulative function  
 RDF - Rayleigh density function  
 RDISCRIM - robust discriminant  
 RDIST - robust distance  
 RDVER - vertical direction for the search radius  
 REPAR - reparametrize  
 REPS - replicates  
 RESID - residuals  
 RIF - Rayleigh inverse function  
 RJS - rejection sampling  
 RMS - root mean square  
 RMSEA - root mean square error of approximation  
 RMSSTD - root mean square standard deviation  
 ROC - receiver operating characteristic  
 ROWPCT - Row percentages  
 RRN - Rayleigh random number  
 RS - response surface  
 RSE - robust standard errors  
 RSEED - random seed  
 RSM - response surface methods  
 RSQ - stress and squared correlation  
 RSS - residual sum of squares  
 RSTATISTICS - row statistics  
 RTF - rich text format  
 RWM-H - random walk Metropolis-Hastings  
 RWSTEM - stem and leaf plot for rows

## S

S chart - standard deviation control chart  
 SANG1 - angle (in degrees) of the first minor axis of the search ellipsoid  
 SANG2 - angle (in degrees) of the major axis of the search ellipsoid  
 SANG3 - angle (in degrees) of the second minor axis of the search ellipsoid  
 SAV - SPSS files  
 SB - smaller the better  
 sc - scale  
 SC - set correlation

SCDFUNC - standardized coefficients for canonical variables  
 SCF - Studentized cumulative function  
 SD - standard deviations  
 sd2/sas7bdat - SAS v9 files  
 SDF - Studentized density function  
 SE/sc/S.E. - standard error  
 SEK - standard error of kurtosis  
 SEM - standard error of mean  
 SES - standard error of skewness  
 shp - shape  
 SIF - Studentized inverse function  
 SIMPLS - Straight-forward Implementation of Partial Least Squares  
 SKMEAN - simple kriging mean  
 SL - specification limit  
 SMIN - minimum split value  
 SPLOM - scatter plot matrix  
 SQL - structured query language  
 SQRT/SQR - square-root  
 SRN - Studentized random number  
 SRWR - sum of rank weighted residuals  
 SS - sum of squares  
 SSCP - sum of squares and cross products  
 STA - Statistica v5 data files  
 STAND - standardized deviates  
 SVD - singular value decomposition  
 SW - Shapiro-Wilks  
 SYC/CMD - SYSTAT command Files  
 SYZ/SYD/SYS - SYSTAT data files  
 SYO - SYSTAT output files

## T

T1 - one-sample t-test  
 T2 - two-sample t-test  
 TANALYZE - Taguchi design: analyze  
 TCF - t cumulative function  
 TCOR - total correlation  
 TCOV - total covariance  
 TDF - t density function  
 TESTAT - Test Item Analysis

TESTATCL - classical test item analysis  
 TESTATLOG - logistic item response analysis  
 TETRA - tetrachoric correlations  
 TGENERATE - Taguchi design: generate  
 TIF - t inverse function  
 TIFF - Tagged Image File Format  
 TLOG - log time  
 TLOSS - Taguchi's Loss Function  
 TNH - hyperbolic tangent  
 TOHC0 - Hypothesis Testing: Zero correlation  
 TOHC1 - Hypothesis Testing: Specific correlation  
 TOHC2 - Hypothesis Testing: Equality of two correlation coefficients  
 TOHP1 - Hypothesis Testing: Single proportion  
 TOHP2 - Hypothesis Testing: Equality of two proportions  
 TOHT1 - Hypothesis Testing: One sample t-test  
 TOHT2 - Hypothesis Testing: Two sample t-test  
 TOHTPAIRED - Hypothesis Testing: Paired t-test  
 TOHV1 - Hypothesis Testing: Single variance  
 TOHV2 - Hypothesis Testing: Two variances  
 TOHVN - Hypothesis Testing: Several variances  
 TOHZ1 - Hypothesis Testing: One sample z-test  
 TOHZ2 - Hypothesis Testing: Two sample z-test  
 TOL - tolerance  
 TPLOT - time series plot  
 TPREDICT - Taguchi design: predict  
 TRCF - Triangular cumulative function  
 TRDF - Triangular density function  
 TRI - triangular  
 TRIF - Triangular inverse function  
 TRIM - trimmed mean  
 TRN - t random number  
 TRP - transpose  
 TRRN - Triangular random number  
 TSFOURIER - Fourier decomposition of time series  
 TSIV - Two-Stage Instrumental Variables  
 TSLS - Two-Stage Least Squares

TSP - traveling salesman path  
 TSQ chart - Hotelling's  $T^2$  chart  
 TSSMOOTH - smoothing time series  
 TXT - text format

## U

U chart - chart showing defects per unit  
 UCF - Uniform cumulative function  
 UCL - upper control limit  
 UDF - Uniform density function  
 UIF - Uniform inverse function  
 UNCE - uncertainty coefficient  
 URN - Uniform random number  
 USL - upper specification limit  
 UTL - upper tolerance limit

## V

VAR - variance  
 VIF - variance inflation factor

## W

WB - Weibull  
 WCF - Weibull cumulative function  
 WCOR - pooled within-group correlation  
 WCOV - pooled within-group covariance  
 WDF - Weibull density function  
 WHISKER - Box-and-Whisker plot  
 WIF - Weibull inverse function  
 WMF - Windows metafile  
 WRN - Weibull random number

## X

XCF - Chi-square cumulative function  
 XDF - Chi-square density function  
 XIF - Chi-square inverse function  
 XLAG - separation distance between lags  
 XLS - excel format  
 XLTOL - tolerance for lags  
 XMAX - maximum along x axis  
 XMIN - minimum along x axis

X-MR chart - Individuals and moving range chart  
XPT/TPT - SAS transport files  
XRN - Chi-square random number  
XTAB - Crosstabulations

## Y

YMAX - maximum along y axis  
YMIN - minimum along y axis

## Z

Z1 - one-sample z-test  
Z2 - two-sample z-test  
ZCF - Normal cumulative function  
ZDF - Normal density function  
ZICF - Zipf cumulative function  
ZIDF - Zipf density function  
ZIF - Normal inverse function  
ZIIF - Zipf inverse function  
ZIRN - Zipf random number  
ZMAX - maximum along z axis  
ZMIN - minimum along z axis  
ZRN - Normal random number



## A

- A matrix, II-192
- accelerated failure time distribution, IV-433
- ACF plots, IV-529
- additive trees, I-80, I-91
- AIC and Schwarz's BIC, II-39, II-108, II-292, II-300, II-344, II-385, III-1, III-258, IV-99, IV-427
  - see linear models, II-17
- Akaike Information Criterion, III-458
- alpha level, IV-22, IV-28
- alternative hypothesis, I-13, IV-20
- analysis of covariance, II-153, II-209
  - examples, II-170
- analysis of variance, II-107
  - AIC and Schwarz's BIC, II-108
  - algorithms, II-171
  - assumptions, II-25
  - between-group differences, II-32
  - commands, II-121
  - compared to loglinear modeling, III-95
  - compared to regression trees, I-45
  - contrasts, II-28, II-113, II-115, II-116
  - data format, II-121
  - examples, II-122, II-126, II-132, II-145, II-146, II-148, II-151, II-155, II-160, II-163, II-166, II-170
  - factorial, II-24
  - homogeneity tests, II-113
  - hypothesis tests, II-23, II-113, II-115, II-116
  - interactions, II-25
  - normality tests, II-112
  - pairwise comparisons, II-117
  - power analysis, IV-19, IV-26, IV-55, IV-57, IV-77, IV-80
  - Quick Graphs, II-121
  - repeated measures, II-31, II-110
  - resampling, II-108
  - residuals, II-110
  - sums of squares, II-113
  - two-way ANOVA, IV-26, IV-57, IV-80
  - unbalanced designs, II-29
  - unequal variances, II-26
  - usage, II-121
  - within-subject differences, II-32
- Anderberg dichotomy coefficients, I-164, I-173
- Anderberg's binary similarity coefficient, I-164
- Anderson-Darling test, I-303
- Andrews procedure, III-279
- angle tolerance, IV-388
- anisotropy, IV-392, IV-405
  - geometric, IV-392
  - zonal, IV-393
- A-optimality, I-364
- ARIMA models, IV-514, IV-523, IV-540
  - algorithms, IV-578
- arithmetic mean, I-299, I-308
- ARMA models, IV-519
- asymptotically distribution-free estimates, III-412
- autocorrelation plots, II-11, IV-516, IV-520
- Automatic Interaction Detection(AID), I-45, I-47
- autoregressive models, IV-516
- average run length curves, IV-134
  - chart types, IV-137
  - continuous distributions, IV-139
  - discrete distributions, IV-140
  - overview, IV-134
  - probability limits, IV-137
- axial designs, I-360



## B

backward elimination, II-15  
 bandwidth, IV-350, IV-355, IV-388  
   optimal values, IV-356  
   relationship with kernel function, IV-357  
 basic statistics  
   Anderson-Darling test, I-303, I-309  
   columns, I-307  
   commands, I-322  
   Cronbach's alpha, I-321  
   examples, I-324, I-326, I-327, I-328, I-333, I-338, I-340, I-341, I-342  
   geometric mean, I-300, I-308  
   harmonic mean, I-300, I-308  
   multivariate normality assessment, I-303  
   N-&P-tiles, I-309  
   overview, I-297  
   Quick Graphs, I-323  
   resampling, I-298  
   rows, I-316  
   Shapiro-Wilk test, I-302, I-309  
   stem-and-leaf for columns, I-314  
   stem-and-leaf for rows, I-320  
   test for normality, I-302  
   trimmed mean, I-299, I-308  
   usage, I-323  
 bayesian regression, II-50  
   credibility intervals, II-50  
   gamma prior, II-52  
   normal prior, II-52  
 best linear unbiased estimates (BLUE), II-344, II-386  
 best linear unbiased predictors (BLUP), II-344, II-386  
 beta level, IV-22  
 between-group differences  
   in analysis of variance, II-32  
 bias, II-15  
 binary logit, III-2  
   compared to multinomial logit, III-5  
 binary trees, I-43  
 biplots, IV-6, IV-8

bisquare procedure, III-279  
 biweight kernel, IV-365  
 Bonferroni inequality, I-47  
 Bonferroni test, I-175, II-27, II-118, II-196, II-307, II-394  
 bootstrap, I-19, I-21  
 box plot, I-305  
 Box-and-Whisker plots, IV-112  
 Box-Behnken designs, I-357, I-380  
 Box-Cox power transformation, IV-157  
 Box-Hunter designs, I-353, I-373  
 Bray-Curtis measure, I-162, I-172  
 broad inference space, II-280  
 C  
 c charts, IV-131  
 C matrix, II-193  
 candidate sets  
   for optimal designs, I-363  
 canonical correlation analysis  
   data format, IV-304  
   examples, IV-305, IV-308, IV-312  
   interactions, IV-304  
   model, IV-299  
   nominal scales, IV-304  
   overview, IV-291  
   partialled variables, IV-300  
   Quick Graphs, IV-305  
   resampling, IV-291  
   rotation, IV-303  
   usage, IV-304  
 canonical rotation, IV-7  
 categorical data, III-321  
 categorical predictors, I-45  
 Cauchy kernel, IV-365  
 CCF plots, IV-531  
 central composite designs, I-356, I-384  
 centroid designs, I-359  
 CHAID, I-46, I-47  
 chi-square tests for independence, I-229, I-233, I-242  
 circle model

- in perceptual mapping, IV-5
- city-block distance, I-172, III-191
- classical analysis, IV-488
- classification and regression trees, I-41
- classification functions, I-396
- classification trees
  - algorithms, I-62
  - basic tree model, I-42
  - commands, I-54
  - compared to discriminant analysis, I-46, I-49, I-46
  - data format, I-54
  - displays, I-51
  - examples, I-55, I-57, I-59
  - loss functions, I-51
  - missing data, I-62
  - mobiles, I-41
  - model, I-51
  - overview, I-41
  - pruning, I-47
  - Quick Graphs, I-54
  - resampling, I-41
  - saving files, I-54
  - stopping criteria, I-47, I-53
  - usage, I-54
- cluster analysis
  - additive trees, I-91
  - algorithms, I-122
  - clustering, I-65
  - commands, I-93
  - data types, I-95
  - distances, I-84
  - examples, I-96, I-105, I-108, I-109, I-111, I-112, I-115, I-116, I-118, I-120
  - exclusive clusters, I-66
  - hierarchical clustering, I-82
  - k-means clustering, I-78
  - k-medians clustering, I-79
  - missing values, I-122
  - overlapping clusters, I-66
  - overview, I-65
  - Quick Graphs, I-95
  - resampling, I-66
  - saving files, I-95
  - usage, I-95
- clustered data, II-421
- clustering
  - hierarchical clustering, I-68
  - k-clustering, I-78
  - validity, I-87
- Cochran's test of linear trend, I-234
- coefficient of alienation, III-190, III-212
- coefficient of determination
  - see multiple correlation
- coefficient of variation, I-307
- Cohen's kappa, I-226, I-234
- communalities, I-458
- compound symmetry, II-32
- conditional logistic regression, III-5
- confidence curves, III-273
- confidence intervals, I-11, I-307
- path analysis, III-455
- conjoint analysis
  - additive tables, I-126
  - algorithms, I-152
  - commands, I-135
  - compared to logistic regression, I-132
  - data format, I-135
  - examples, I-136, I-140, I-143, I-147
  - missing data, I-153
  - model, I-133
  - multiplicative tables, I-128
  - overview, I-125
  - Quick Graphs, I-135
  - resampling, I-125
  - saving files, I-135
  - usage, I-135
- constraints
  - in mixture designs, I-360
- contingency coefficient, I-227
- contour plot, IV-243
- contour plots, IV-401
- contrast coefficients, II-31
- contrasts
  - in analysis of variance, II-28
- control charts

- aggregated data, IV-120
- average run length curves, IV-136
- control limits, IV-121
- discrete control limits, IV-121
- operating characteristic curves, IV-135
- raw data, IV-120
- regression charts, IV-152
- sigma limits, IV-122
- convergence, III-98
- convex hulls, IV-398
- Cook's distance, II-12
- Cook-Weisberg graphical confidence curves, III-273
- coordinate exchange method, I-363, I-386
- correlations, I-67, I-157
  - algorithms, I-199
  - binary data, I-173
  - canonical, IV-291
  - commands, I-177
  - continuous data, I-171
  - data format, I-178
  - dissimilarity measures, I-172
  - distance measures, I-172
  - examples, I-179, I-182, I-185, I-186, I-188, I-192, I-195, I-196, I-198
  - missing values, I-170, I-199, III-135
  - options, I-174
  - overview, I-157
  - power analysis, IV-19, IV-25, IV-42, IV-44
  - Quick Graphs, I-178
  - rank-order data, I-172
  - resampling, I-158
  - saving files, I-179
  - set, IV-291
  - usage, I-178
- correlograms, IV-403
- correspondence analysis, IV-2, IV-6
  - algorithms, I-218
  - commands, I-206
  - data format, I-206
  - examples, I-207, I-214
  - missing data, I-218
  - model, I-204
  - overview, I-201
  - Quick Graphs, I-206
  - resampling, I-201
  - simple correspondence analysis, I-204
  - usage, I-206
- covariance matrix, I-171, III-135
- covariance paths
  - path analysis, III-401
- covariograms, IV-387
- Cox-Snell residual plot, IV-434
- Cramer's V, I-227
- critical level, I-13
- Cronbach's alpha, IV-488, IV-489
  - see basic statistics, I-321
- crossover designs, II-175
- crosstabulation
  - commands, I-244
  - data format, I-246
  - examples, I-248, I-250, I-253, I-256, I-257, I-258, I-261, I-263, I-269, I-271, I-273, I-275, I-277, I-279, I-293
  - multiway, I-237
  - one-way, I-220, I-222, I-228
  - overview, I-219
  - Quick Graphs, I-247
  - resampling, I-219
  - standardizing tables, I-221
  - two-way, I-220, I-223, I-231
  - usage, I-246
- cross-validation, I-48, I-396, II-16, III-360
- cumulative sum charts
  - see cusum charts, IV-142
- D
  - D matrix, II-194, II-288, II-309, II-355, II-397
  - D SUB-A ( $d_a$ ), IV-321
  - dates, IV-430
  - dendrograms, I-65, I-107
  - dependence paths
    - path analysis, III-399
  - descriptive statistics, I-1
    - see basic statistics, I-307



- design of experiments, I-132, I-368, I-369
  - axial designs, I-360
  - Box-Behnken designs, I-357
  - central composite designs, I-356
  - centroid designs, I-359
  - commands, I-370
  - examples, I-371, I-372, I-373, I-375, I-377, I-379, I-380, I-381, I-382, I-384, I-386
  - factorial designs, I-349, I-350
  - lattice designs, I-359
  - mixture designs, I-350, I-357
  - optimal designs, I-350, I-362
  - overview, I-345
  - Quick Graphs, I-371
  - response surface designs, I-350, I-354
  - screening designs, I-360
  - usage, I-370
- determinant criterion
  - see D-optimality
- Dice's binary similarity coefficient, I-164
- dichotomy coefficients, I-164
  - Anderberg, I-173
  - Jaccard, I-173
  - positive matching, I-173
  - simple matching, I-173
  - Tanimoto, I-173
- difficulty, IV-507
- discrete choice model, III-7
  - compared to polytomous logit, III-8
- discrete gaussian convolution, IV-361
- discriminant analysis
  - classical discriminant analysis, I-400
  - commands, I-407
  - data format, I-408
  - estimation, I-401
  - examples, I-409, I-413, I-420, I-427, I-435, I-438, I-444, I-449
  - linear discriminant function, I-397
  - model, I-400
  - multiple groups, I-399
  - options, I-401
  - overview, I-391
  - prior probabilities, I-398
  - Quick Graphs, I-408
  - resampling, I-391
  - robust discriminant analysis, I-399
  - statistics, I-404
  - stepwise estimation, I-401
  - usage, I-408
- discrimination parameter, IV-507
- dissimilarities
  - direct, III-187
  - indirect, III-187
- distance measures, I-67, I-157
- distances
  - nearest-neighbor, IV-396
- distance-weighted least squares (DWLS) smoother, IV-361
- distributions
  - Benford's law, I-499, III-332, IV-86, IV-221
  - beta, I-500, III-333, III-335, IV-88, IV-222
  - binomial, I-499, III-332, IV-86, IV-221
  - Cauchy, I-500, III-333, III-335, IV-88, IV-222
  - chi-square, I-500, III-333, III-335, IV-88, IV-222
  - discrete uniform, I-499, III-332, IV-86, IV-221
  - double exponential, I-501, III-335, IV-88, IV-222
  - Erlang, I-501, III-335, IV-88, IV-222
  - exponential, I-501, III-333, III-336, IV-88, IV-222
  - F, III-333, III-336, IV-88, IV-222
  - gamma, I-501, III-333, III-336, IV-89, IV-222
  - generalized lambda, IV-222
  - geometric, I-499, III-332, IV-86, IV-221
  - Gompertz, I-501, III-333, III-336, IV-89, IV-222
  - Gumbel, I-501, III-333, III-336, IV-89, IV-222
  - hypergeometric, I-499, III-332, IV-86, IV-221
  - inverse Gaussian, I-501, III-333, III-336, IV-89, IV-222
  - logarithmic series, I-499, III-332, IV-87, IV-221
  - logistic, I-501, III-333, III-336, IV-89, IV-222



- logit normal, I-501, III-333, III-336, IV-89, IV-222
- loglogistic, I-501, III-333, III-336, IV-89, IV-222
- lognormal, I-501, III-333, III-336, IV-89, IV-222
- negative binomial, I-499, III-333, IV-87, IV-221
- non-central chi-square, III-333, III-336, IV-89, IV-222
- non-central F, III-333, III-336, IV-89, IV-222
- non-central t, III-333, III-336, IV-89, IV-222
- normal, I-501, III-333, III-336, IV-89, IV-222
- Pareto, I-501, III-333, III-336, IV-89, IV-222
- Poisson, I-499, III-333, IV-87, IV-221
- Rayleigh, I-501, III-333, III-336, IV-89, IV-223
- smallest extreme value, I-501, III-333, III-336, IV-89, IV-223
- studentized maximum modulus, III-333, III-336, IV-89
- Studentized range, III-336
- studentized range, III-333, IV-89, IV-223
- t, III-333, III-336, IV-89, IV-223
- triangular, I-501, III-334, III-336, IV-89, IV-223
- uniform, I-501, III-334, III-336, IV-89, IV-223
- Weibull, I-501, III-334, III-336, IV-89, IV-223
- zipf, I-499, III-333, IV-87, IV-221
- dit plots, I-14
- D-optimality, I-364
- dot histogram plots, I-14
- Double, III-333
- D-Prime ( $d'$ ), IV-320
- dummy codes, II-180
- Duncan test, II-27, II-119, II-197
- Dunnett test, II-27, II-119, II-197
- Dunnett's T3 test, II-27, II-119, II-197
- Dunn-Sidak test, I-175
- E
- ECVI, III-458
- edge effects, IV-398
- effect size
  - in power analysis, IV-22, IV-23
- effects coding, II-20, II-180
- efficiency, I-362
- eigenvalues, I-405
- ellipse model
  - in perceptual mapping, IV-6
- EM algorithm, I-492
- EM estimation, III-130
  - for correlations, I-175, III-135
  - for covariance, III-135
  - for SSCP matrix, III-135
- endogenous variables
  - path analysis, III-400
- Epanechnikov kernel, IV-364
- equamax rotation, I-460, I-464
- Erlang, III-333
- Estimation, III-135
- Euclidean distances, III-188
- exogenous variables
  - path analysis, III-400
- expected cross-validation index, III-458
- Exponential, III-336
- exponential distribution, IV-432
- exponential model, IV-390, IV-404
- exponential smoothing, IV-524
- exponentially weighted moving average charts, IV-146
  - control limits, IV-147
- external unfolding, IV-4
- F
- F, III-333
- F and R matrices, II-308, II-354, II-396
- F distribution
- F matrix, II-287
- factor analysis, I-457, IV-2
  - algorithms, I-492
  - commands, I-468

- compared to principal components analysis, I-460
- convergence, I-463
- correlations vs covariances, I-457
- eigenvalues, I-463
- eigenvectors, I-467
- examples, I-469, I-473, I-476, I-478, I-482, I-485
- iterated principal axis, I-463
- loadings, I-467
- maximum likelihood, I-463
- missing values, I-492
- number of factors, I-463
- overview, I-453
- principal components, I-463
- Quick Graphs, I-468
- resampling, I-453
- residuals, I-465
- rotation, I-459, I-464
- save, I-466
- scores, I-466
- usage, I-468
- factor loadings, IV-488
- factorial analysis of variance, II-24
- factorial designs, I-349, I-350
  - analysis of, I-353
  - examples, I-371
  - fractional factorials, I-352
  - full factorial designs, I-352
- F-distribution
  - non-centrality parameter, IV-60
- Fedorov method, I-363
- Fieller bounds, III-48
- filters, IV-527
- Fisher's exact test, I-226, I-233
- Fisher's linear discriminant function, IV-2
- Fisher's LSD, II-197
- Fisher's LSD test, II-27, II-118, II-307, II-395
- fitting distributions
  - commands, I-501
  - examples, I-504, I-505, I-507, I-508, I-510, I-511, I-513
  - goodness-of-fit tests, I-496
  - maximum likelihood method, I-497
  - method of moments, I-497
  - method of quantiles or order statistic, I-497
  - overview, I-495
  - Quick Graphs, I-503
  - Shapiro-Wilk's test for normality, I-497
  - usage, I-503
- fixed effects, II-279
- fixed variance
  - path analysis, III-402
- fixed-bandwidth method
  - compared to KNN method, IV-357
  - for smoothing, IV-355, IV-357, IV-364
- Fletcher-Powell minimization, IV-507
- forward selection, II-15
- Fourier analysis, IV-526, IV-545
- fractional factorial designs
  - Box-Hunter designs, I-353
  - examples, I-372, I-373, I-375, I-377, I-379
  - homogeneous fractional designs, I-353
  - Latin square designs, I-353
  - mixed-level fractional designs, I-353
  - Plackett-Burman designs, I-353
  - Taguchi designs, I-353
- Freeman-Tukey deviates, III-93, III-102
- frequencies, I-23, I-54, I-135, I-179, I-206, I-246, I-248, I-323, I-408, I-468, I-469, I-503, I-544, II-54, II-121, II-122, II-202, II-310, II-357, II-399, II-441, III-23, III-103, III-104, III-137, III-194, III-217, III-283, III-339, III-364, III-385, III-413, IV-9, IV-62, IV-63, IV-103, IV-162, IV-244, IV-280, IV-305, IV-325, IV-328, IV-366, IV-410, IV-449, IV-495, IV-498, IV-547, IV-587
- frequency tables, III-93, III-102
  - see crosstabulation
- Friedman test, III-328
- G
  - Gabriel test, II-27, II-119, II-197
  - Games-Howell test, II-27, II-119, II-197
  - Gaussian kernel, IV-364, IV-365
  - Gaussian model, IV-390, IV-404

Gauss-Newton method, III-269, III-272  
 general linear models, II-175  
   algorithms, II-249  
   categorical variables, II-179  
   commands, II-200  
   contrasts, II-189, II-191  
   data format, II-201  
   examples, II-203, II-211, II-212, II-213, II-215, II-217, II-220, II-222, II-224, II-234, II-237, II-238, II-242, II-246, II-247, II-248  
   hypothesis options, II-188  
   hypothesis tests, II-186  
   mixture model, II-184  
   model estimation, II-177  
   overview, II-175  
   pairwise comparisons, II-195  
   post hoc tests, II-199  
   Quick Graphs, II-202  
   resampling, II-176  
   stepwise regression, II-183  
   usage, II-201  
 generalized least squares, III-412, IV-584  
 generalized variance, IV-294  
 geometric mean, I-300, I-308  
 geostatistical models, IV-386, IV-387  
 between-groups testing, III-239  
 Gini index, I-48, I-51  
 GLM  
   see general linear models, II-175  
 global criterion  
   see G-optimality  
 GMA chart, IV-146  
 Goodman-Kruskal gamma, I-227, I-234  
 Goodman-Kruskal lambda, I-234  
 goodness-of-fit tests, I-496  
 G-optimality, I-364  
 Gower2 binary similarity coefficient, I-164  
 Graeco-Latin square designs, I-353  
 Greenhouse-Geisser statistic, II-33  
 Guttman  $\mu_2$  monotonicity coefficients, I-162  
 Guttman's coefficient of alienation, III-190  
 Guttman's loss function, III-212

Guttman-Rulon coefficient, IV-489

## H

Hadi outlier detection, I-168  
 Hamman's binary similarity coefficient, I-164  
 Hampel procedure, III-279  
 Hanning weights, IV-512  
 harmonic mean, I-300, I-308  
 hazard function  
   heterogeneity, IV-435  
 Henderson's mixed model equations, II-279, II-293  
 Henze-Zirkler test, I-303  
 heteroskedasticity, IV-583  
 heteroskedasticity-consistent standard errors, IV-583  
 hierarchical clustering, I-68, I-82  
   distances, I-84  
   validity index, I-75  
 hierarchical linear mixed models  
   categorical variables, II-389  
   commands, II-398  
   examples, II-399, II-402, II-406, II-408, II-412, II-414, II-417  
   hypothesis testing, II-394  
   model estimation, II-387  
   options, II-392  
   overview, II-385  
   Quick Graphs, II-398  
   random effects, II-390  
   usage, II-398  
 hierarchical linear models  
   see mixed regression  
 hinge, I-301  
 Hochberg's GT2 test, II-27, II-119, II-197, II-307, II-395  
 hole model, IV-391, IV-405  
 Holt's method, IV-524  
 homogeneity tests, II-113  
   Levene's test, II-113  
 Hotelling's T squared charts, IV-153  
 Hotelling-Lawley trace, III-226  
 Huber procedure, III-279



Huynh-Feldt statistic, II-33

hyper-Graeco-Latin square designs, I-353

hypothesis

alternative, I-13

null, I-13

testing, I-12, II-7

hypothesis testing

Bartlett's test, I-521

commands, I-541

confidence intervals, I-520, I-521, I-522

data format, I-543

examples, I-544, I-545, I-547, I-548, I-549, I-551, I-552, I-556, I-557, I-560, I-562, I-564

Levene's tests, I-521

multiple tests, I-522

overview, I-519

Quick Graphs, I-544

resampling, I-519

test for means, I-520

tests for correlation, I-522

tests for mean, I-520

tests for proportion, I-520, I-538

tests for variance, I-521

usage, I-543

## I

ID3, I-47

I-MR chart

see X-MR chart, IV-150

incomplete block designs, II-175

independence, I-223

in loglinear models, III-94

individual cases charts

See X charts, IV-129

INDSCAL model, III-185

inertia, I-202

inferential statistics, I-7, IV-20

instrumental variables, IV-582

intermediate inference space, II-280

internal-consistency, IV-489

interquartile range, I-301

interval censored data, IV-428

inverse-distance smoother, IV-360

isotropic, IV-387

item-response analysis

see test item analysis

item-test correlations, IV-488

## J

Jaccard dichotomy coefficients, I-164, I-173

jackknife, I-18, I-22

jackknifed classification matrix, I-396

## K

k nearest-neighbors method

compared to fixed-bandwidth method, IV-357

for smoothing, IV-356, IV-362

k-clustering, I-78

k-means, I-78

k-medians, I-79

Kendall's Tau b, I-172

Kendall's tau-b coefficient, I-227

kernel functions, IV-350, IV-352

biweight, IV-364

Cauchy, IV-364

Epanechnikov, IV-364

Gaussian, IV-364

plotting, IV-354

relationship with bandwidth, IV-357

tricube, IV-364

triweight, IV-362, IV-364

k-exchange method, I-363

Kolmogorov-Smirnov test, III-319

KR20, IV-489

kriging, IV-405

ordinary, IV-394, IV-405, IV-407

simple, IV-393, IV-407

trend components, IV-394

universal, IV-394, IV-407

Kruskal's loss function, III-211

Kruskal's STRESS, III-190

Kruskal-Wallis test, III-319

K-S test, III-319



Kulczynski's binary similarity coefficient, I-164  
 Kulczynski's binary similarity coefficient, I-173  
 kurtosis, I-307

## L

latent trait model, IV-488, IV-490  
 Latin square designs, I-353, I-375  
 lattice, III-382  
 lattice designs, I-359  
 least absolute deviations, III-268  
 least absolute deviations regression, IV-260  
 least median of squares regression, IV-261  
   search method, IV-269  
 least trimmed squares regression, IV-261  
 Levene test, II-25  
 leverage, II-12  
 likelihood ratio chi-square, I-233, III-96, III-101  
   compared to Pearson chi-square, III-96  
 likelihood-ratio chi-square, I-226  
 Lilliefors test, III-334, III-355  
 linear contrasts, II-28  
 linear discriminant model, I-392  
 linear mixed models  
   categorical variables, II-347  
   commands, II-356  
   examples, II-357, II-362, II-366, II-369, II-372, II-379, II-382  
   hypothesis testing, II-352  
   model estimation, II-345  
   options, II-350  
   overview  
   Quick Graphs, II-356  
   random effects, II-348  
   usage, II-356  
 linear models  
   general linear models, II-175  
   hierarchical, II-421  
   linear discriminant model, I-392  
   linear regression, II-39, II-299, II-385  
 linear regression, I-11, II-7, II-39  
   AIC and Schwarz's BIC, II-39  
   Anderson-Darling test, II-45

  bayesian, II-50  
   commands, II-53  
   data format, II-54  
   examples, II-55, II-60, II-63, II-67, II-71, II-75, II-81, II-83, II-85, II-86, II-87, II-89, II-95, II-97, II-99  
 Kolmogorov-Smirnov test, II-45  
 model, II-41  
 normality tests, II-45  
 overview, II-39  
 prediction intervals, II-40, II-46  
 Quick Graphs, II-54  
 resampling, II-40, II-47  
 residuals, II-9, II-41  
 ridge, II-48  
 Shapiro-Wilk test, II-45  
 stepwise, II-15  
 tolerance, II-43  
 usage, II-54  
   using correlation matrix as input, II-18, II-89  
   using covariance matrix as input, II-18, II-89  
   using SSCP matrix as input, II-18, II-89  
   variance inflation factor, II-70  
 listwise deletion, I-492, III-125  
 Little's MCAR test, III-123, III-133  
 loadings, I-456, I-457  
 LOESS smoothing, IV-361, IV-363, IV-367, IV-368, IV-370, IV-380  
 logistic item-response analysis, IV-506  
   one-parameter model, IV-490  
   two-parameter model, IV-490  
 logistic regression  
   AIC and Schwarz's BIC, III-1  
   algorithms, III-85  
   categorical predictors, III-11  
   classification table, III-17  
   compared to conjoint analysis, I-132  
   conditional variables, III-10  
   confidence intervals, III-48  
   data format, III-22  
   deciles of risk, III-17  
   discrete choice, III-13  
   dummy coding, III-11, III-12

- effect coding, III-11, III-12
- estimation, III-15
- examples, III-24, III-27, III-33, III-39, III-45, III-50, III-60, III-69, III-70, III-77, III-81
- missing data, III-86
- model, III-10
- options, III-14
- overview, III-1
- post hoc tests, III-20
- prediction table, III-16
- quantiles, III-18, III-49
- Quick Graphs, III-23
- regression diagnostics, III-87
- robust standard errors, III-16
- ROC curve, III-1
- simulation, III-19
- usage, III-22
- weights, III-23
- logit
  - binary logit, III-2
  - conditional logit, III-5
  - discrete choice logit, III-7
  - multinomial logit, III-5
  - stepwise logit, III-9
- loglinear modeling
  - commands, III-103
  - compared to analysis of variance, III-95
  - compared to Crosstabs, III-102
  - convergence, III-96
  - data format, III-103
  - examples, III-105, III-114, III-117, III-121
  - frequency tables, III-102
  - model, III-96
  - overview, III-93
  - parameters, III-100
  - Quick Graphs, III-104
  - saturated models, III-95
  - statistics, III-100
  - structural zeros, III-98
  - usage, III-103
- log-logistic distribution, IV-432
- lognormal distribution, IV-432
- longitudinal data, II-421
- loss function, III-265
  - multidimensional scaling, III-210
- loss functions, I-48
- LOWESS smoothing, IV-513
- low-pass filter, IV-527
- LSD test, II-197
- M
  - madograms, IV-403
  - Mahalanobis distances, I-392
  - Mann-Whitney, III-342
  - Mantel-Haenszel test, I-238
  - Mardia skewness and kurtosis, I-298, I-303
  - Marquardt method, III-275
  - Marron & Nolan canonical kernel width, IV-357, IV-364
  - mass, I-202
  - matrix displays, I-70
  - maximum likelihood estimates, II-385, III-266
  - maximum likelihood factor analysis, I-461
  - Maximum Wishart likelihood, III-411
  - McFadden's conditional logit model, III-7
  - McNemar's test, I-226, I-234
  - MDPREF, IV-6, IV-8
  - MDS
    - see multidimensional scaling, III-185
  - mean, I-3, I-307
  - mean smoothing, IV-358, IV-365
  - means coding, II-21
  - median, I-4, I-299, I-307
  - median smoothing, IV-358
  - meta-analysis, II-19
  - midrange, I-301
  - minimum spanning trees, IV-396
  - Minkowski metric, III-191
  - MIS function, III-142
  - Missing At Random(MAR), III-131
  - Missing Completely At Random(MCAR), III-131
  - missing value analysis
    - casewise pattern table, III-142
    - data format, III-137

- EM algorithm, III-130, III-134, III-135, III-154, III-168, III-176
- examples, III-137, III-142, III-154, III-168, III-176
- listwise deletion, III-125, III-154, III-168
- MISSING command, III-136
- missing value patterns, III-137
- model, III-134
- outliers, III-135
- overview, III-123
- pairwise deletion, III-125, III-154, III-168
- pattern variables, III-124, III-176
- Quick Graphs, III-137
- randomness, III-131
- regression imputation, III-127, III-134, III-154, III-176
- resampling, III-123
- saving estimates, III-134, III-137
- unconditional mean imputation, III-126
- usage, III-137
- mixed models, II-251
  - AIC and Schwarz's BIC, II-292
  - ANOVA Method, II-281
  - compound symmetry structure, II-270
  - covariance structures, II-269
  - diagonal structure, II-271
  - estimation methods, II-281
  - hypothesis testing, II-286
  - MIVQUE(0) method, II-283
  - ML method, II-284
  - pairwise comparison, II-290
  - post hoc tests, II-290
  - REML method, II-285
  - setup, II-267
  - unstructured (general symmetric structure), II-272
  - variance components structure, II-270
- mixed regression
  - algorithms, II-484
  - commands, II-441
  - data format, II-441
  - examples, II-442, II-449, II-457, II-473
  - overview, II-421
  - Quick Graphs, II-441
  - usage, II-441
- mixture designs, I-350, I-357
  - analysis of, I-361
  - axial designs, I-360
  - centroid designs, I-359
  - constraints, I-360
  - examples, I-381, I-382
  - lattice designs, I-359
  - Scheffé model, I-361
  - screening designs, I-360
  - simplex, I-359
- models, I-10, II-301
  - estimation, I-10
- moving average, IV-355, IV-511, IV-517
- moving average chart, IV-144
- moving-averages smoother, IV-360
- M-regression, IV-261
- multidimensional scaling, III-185, IV-2
  - algorithms, III-211
  - assumptions, III-186
  - commands, III-194
  - configuration, III-189, III-193
  - confirmatory, III-193
  - convergence, III-192
  - data format, III-194
  - dissimilarities, III-187
  - distance metric, III-189
  - examples, III-195, III-198, III-200, III-203, III-208
  - Guttman method, III-212
  - individual differences, III-185
  - Kruskal method, III-211
  - log function, III-191
  - loss function, III-190
  - metric, III-189
  - missing values, III-212
  - nonmetric, III-189
  - overview, III-185
  - power function, III-191
  - Quick Graphs, III-194
  - residuals, III-192
  - R-metric, III-191



- Shepard diagrams, III-189, III-194
- usage, III-194
- multilevel models
  - see mixed regression
- multinomial logit, III-5
  - compared to binary logit, III-5
- multinormal tests, III-215
  - examples, III-218, III-219
  - Henze-Zirkler test, III-215
  - Mardia skewness and kurtosis, III-215
  - overview, III-215
  - Quick Graphs, III-217
  - usage, III-217
  - using commands, III-217
- multiple comparison tests
  - see pairwise comparisons, II-117, II-195
- multiple correlation, II-8
- multiple correspondence analysis, I-203
- multiple regression, II-12
- multiple tests
  - Bonferroni adjustment, I-522
  - Dunn-Sidak adjustment, I-522
- multivariate analysis of variance, III-223
  - between-groups testing, III-239
  - categorical variables, III-229
  - commands, III-244
  - data format, III-244
  - examples, III-246, III-248, III-253, III-255, III-257, III-258
  - Hotelling-Lawley trace, III-226
  - hypothesis test, III-232
  - overview, III-223
  - Pillai trace, III-225
  - post hoc test, III-242
  - Quick Graphs, III-245
  - repeated measures, III-230
  - Roy's Greatest root, III-226
  - usage, III-244
  - Wilks' lambda, III-225
  - within-group testing, III-241
- multivariate normality assessment
  - Henze-Zirkler test, I-303
  - Mardia's skewness, I-303
- mutually exclusive, I-222
- N
- N- & P-tiles, I-309
  - methods, I-311
  - transformation, I-309
- Nadaraya-Watson smoother, IV-360
- narrow inference space, II-280
- Nelson-Aalen cumulative hazard estimator, IV-438
- nesting, II-175
- Newton-Raphson method, III-93
- NIPALS (Nonlinear Iterative Partial Least Squares)
  - see partial least squares regression, III-377
- nodes, I-43
- nominal data, III-321
- non-central F-distribution, IV-34, IV-60
- non-centrality parameters, IV-34
- nonlinear models, III-261
  - algorithms, III-316
  - commands, III-283
  - computation, III-274, III-316
  - convergence, III-274, III-275
  - data format, III-283
  - estimation, III-269
  - examples, III-284, III-287, III-290, III-293, III-296, III-298, III-299, III-301, III-306, III-311, III-313, III-315
  - functions of parameters, III-277
  - loss functions, III-265, III-270, III-280, III-281
  - missing data, III-316
  - model, III-270
  - parameter bounds, III-274
  - problems, III-269
  - Quick Graphs, III-283
  - recalculation of parameters, III-276
  - resampling, III-261
  - robust estimation, III-278
  - starting values, III-274
  - usage, III-283
- nonmetric unfolding model, III-185
- nonparametric statistics, III-325



## nonparametric tests

- algorithms, III-355
- Anderson-Darling test, III-334
- commands, III-325, III-331, III-338
- data format, III-339
- examples, III-340, III-342, III-343, III-345, III-346, III-347, III-348, III-349, III-350, III-353, III-354
- Friedman test, III-328
- independent samples test, III-322, III-323
- Kolmogorov-Smirnov test, III-323, III-331
- Kruskal-Wallis test, III-322
- Mann-Whitney test, III-322
- overview, III-319
- Quade test, III-329
- Quick Graphs, III-339
- related variables tests, III-325, III-326, III-328
- resampling, III-319
- sign test, III-325, III-326
- usage, III-339
- Wald-Wolfowitz runs test, III-337
- Wilcoxon Signed-Rank test, III-326

## normal distribution, I-301

## normality tests, II-45, II-112

- Anderson-Darling, II-113
- Anderson-Darling test, II-45
- Kolmogorov-Smirnov test, II-45, II-112
- Shapiro-Wilk, II-112
- Shapiro-Wilk test, II-45

## np charts, IV-129

## NPAR, IV-320

## null hypothesis, I-12, IV-20

## O

## oblimin rotation, I-460, I-464

## observational studies, I-347

## OC curves, IV-134

## Occam's razor, I-130

## Ochiai's binary similarity coefficient, I-164

## odds ratio, I-233

## omni-directional variograms, IV-388

## operating characteristic curves

## chart type, IV-136

## continuous distributions, IV-139

## discrete distributions, IV-140

## overview, IV-134

## probability limits, IV-136

## sample size, IV-138

## scaling, IV-138

## optimal designs, I-350, I-362

## analysis of, I-364

## A-optimality, I-364

## candidate sets, I-363

## coordinate exchange method, I-363, I-386

## D-optimality, I-364

## efficiency criteria, I-364

## Fedorov method, I-363

## G-optimality, I-364

## k-exchange method, I-363

## model, I-365

## optimality criteria, I-364

## optimality, I-362

## ORDER, IV-431

## ordinal data, III-320

## Ordinary least squares, III-412

## orthomax rotation, I-460, I-464

## Output, IV-99

## P

## p charts, IV-130

## PACF plots, IV-530

## pairwise comparisons, II-26, II-107, II-117

## Bonferroni test, II-118, II-196

## Duncan test, II-119, II-197

## Dunnett test, II-119, II-197

## Dunnett's T3 test, II-119, II-197

## Fisher's LSD, II-197

## Fisher's LSD test, II-118

## Gabriel test, II-119, II-197

## Games - Howell test, II-197

## Games-Howell test, II-119

## Hochberg's GT2 test, II-119

## Hochberg's test GT2, II-197

## R-E-G-W Q test, II-197

- R-E-G-W-Q test, II-119
- Scheffé test, II-27, II-118, II-197
- Sidak test, II-118, II-197
- Student-Newman-Keuls test, II-119, II-197
- Tamhane's T2 test, II-119, II-197
- Tukey test, II-118, II-196
- Tukey's b test, II-119, II-197
- pairwise deletion, I-492, III-125
- parameters, I-10
- parametric modeling, IV-432
- Pareto charts, IV-111
- partial autocorrelation plots, IV-519, IV-520
- partial least squares regression
  - algorithms, III-377
  - cross-validation, III-363
  - examples, III-365, III-368, III-371, III-375
  - latent factors, III-357, III-359
  - leave-one-out, III-360, III-363
  - NIPALS, III-362
  - PRESS statistic, III-360
  - Quick Graphs, III-364
  - random exclusion, III-360, III-364
  - SIMPLS, III-362
  - test set, III-360
  - training set, III-360
  - usage, III-364
  - using commands, III-364
- partialing
  - in set correlation, IV-295
- partially ordered scalogram analysis with coordinates
  - algorithms, III-395
  - commands, III-385
  - Convergence, III-384
  - convergence, III-384
  - data format, III-385
  - displays, III-383
  - examples, III-386, III-388, III-390
  - missing data, III-395
  - model, III-384
  - overview, III-381
  - Quick Graphs, III-385
  - resampling, III-381
  - usage, III-385
- path analysis
  - algorithms, III-454
  - confidence intervals, III-455
  - covariance paths, III-401
  - covariance relationship, III-409
  - data format, III-413
  - dependence paths, III-399
  - dependence relationship, III-407
  - endogenous variables, III-400
  - estimate, III-411
  - examples, III-414, III-419, III-434, III-442
  - exogenous variables, III-400
  - fixed variance, III-402
  - free parameters, III-418
  - latent variables, III-404
  - manifest variables, III-410
  - measures of fit, III-455
  - method of estimation, III-411
  - model, III-452
  - model statement, III-407
  - options, III-411
  - overview, III-397
  - path diagrams, III-397
  - Quick Graphs, III-413
  - starting values, III-412
  - usage, III-413
  - variance paths, III-401
- Pearson chi-square, I-223, I-228, I-233, III-94, III-101
  - compared to likelihood ratio chi-square, III-96
- Pearson correlation, I-160, I-171
- perceptual mapping
  - algorithms, IV-16
  - commands, IV-9
  - data format, IV-9
  - examples, IV-9, IV-11, IV-12, IV-14
  - methods, IV-8
  - missing data, IV-16
  - model, IV-7
  - overview, IV-1
  - PREFMAP, IV-1
  - Quick Graphs, IV-9

- usage, IV-9
- periodograms, IV-527
- permutation tests, I-222
- phi coefficient, I-48, I-51, I-52, I-227
- Pillai trace, III-225
- Plackett-Burman designs, I-353, I-379
- point processes, IV-386, IV-395
- polynomial contrasts, II-28, II-31, II-192
- polynomial smoothing, IV-358, IV-365
- populations, I-7
- POSET, III-381
- positive matching dichotomy coefficients, I-164, I-173
- Post hoc Test for Repeated measures, III-242
- power, IV-22
- power analysis
  - analysis of variance, IV-19
  - commands, IV-62
  - correlation coefficients, IV-25, IV-42, IV-44
  - correlations, IV-19
  - data format, IV-62
  - examples, IV-63, IV-67, IV-72, IV-77, IV-80
  - generic, IV-34, IV-60, IV-77
  - one-sample t-test, IV-26
  - one-sample z-test, IV-46
  - one-way ANOVA, IV-26, IV-55, IV-77
  - overview, IV-19
  - paired t-test, IV-26, IV-51, IV-67
  - power curves, IV-62
  - proportions, IV-19, IV-25, IV-39, IV-40, IV-63
  - Quick Graphs, IV-62
  - randomized block designs, IV-19
  - t-tests, IV-19
  - two-sample t-test, IV-53, IV-72
  - two-sample z-test, IV-48
  - two-way ANOVA, IV-26, IV-57, IV-80
  - usage, IV-62
  - z-tests, IV-19
- power curves, IV-62
  - overlying curves, IV-67
  - response surfaces, IV-67
- Power model, IV-391, IV-405
- prediction intervals, II-40, II-46
- preference curves, IV-4
- preference mapping, IV-2
- PREFMAP, IV-7
- PRESS statistic
  - in partial least squares regression, III-360
- principal components, I-463
- principal components analysis
  - coefficients, I-456
  - compared to factor analysis, I-460
  - compared to linear regression, I-455
  - loadings, I-456
- prior probabilities, I-398
- probability calculator
  - examples, IV-90, IV-93, IV-94, IV-95
  - overview, IV-85
  - usage, IV-90
- probability limits, IV-121
- probability plots, I-15, II-9
- probit analysis
  - AIC and Schwarz's BIC, IV-99
  - algorithms, IV-107
  - categorical variables, IV-102
  - commands, IV-103
  - data format, IV-103
  - dummy coding, IV-102
  - effect coding, IV-103
  - examples, IV-104, IV-106
  - interpretation, IV-100
  - missing data, IV-107
  - model, IV-100
  - overview, IV-99
  - Quick Graphs, IV-103
  - saving files, IV-103
  - usage, IV-103
- process capability analysis, IV-155
  - Box-Cox power transformation, IV-157
  - non-normal data, IV-157, IV-158
  - process performance, IV-158
- Procrustes rotations, IV-7
- proportional hazards models, IV-433
- proportions
  - power analysis, IV-19, IV-25, IV-39, IV-40,



- IV-63  
p-value, IV-20
- Q**
- QSK**  
coefficients, I-172  
Quade test, III-329  
multiple comparisons, III-329  
pairwise comparisons, III-330  
quadrat counts, IV-385, IV-398  
quadratic contrasts, II-28  
quality analysis, IV-109  
aggregated data, IV-120  
average run length curves, IV-136  
Box-and-Whisker plots, IV-112  
commands, IV-161  
control charts, IV-114  
control limits, IV-121  
cusum charts, IV-142  
data format, IV-162  
discrete control limits, IV-121  
examples, IV-163, IV-164, IV-165, IV-166,  
IV-167, IV-168, IV-176, IV-178, IV-  
180, IV-183, IV-189, IV-191, IV-  
195, IV-197, IV-198, IV-199, IV-  
201, IV-203, IV-204, IV-206, IV-  
207, IV-209, IV-212, IV-213, IV-  
215  
histogram, IV-110  
moving average chart, IV-144  
moving range, IV-149  
operating characteristic curves, IV-135  
overview, IV-109  
Pareto charts, IV-111  
process capability analysis, IV-155  
quick graphs, IV-162  
raw data, IV-120  
regression charts, IV-152  
run charts, IV-114  
run tests, IV-118  
shewhart control charts, IV-116  
sigma limits, IV-122  
TSQ charts, IV-153  
usage, IV-162  
X-MR charts, IV-149
- quantile plots, IV-434  
quantitative symmetric dissimilarity coefficient, I-162  
quartimax rotation, I-460, I-464  
quasi-independence, III-98  
Quasi-Newton method, III-269, III-273
- R**
- R charts, IV-128  
R charts:plotting with X-bar charts, IV-129  
R matrix, II-289  
Ramsay procedure, III-279  
random coefficient models  
see mixed regression  
random effects, II-259, II-390  
in mixed regression, II-421  
random fields, IV-386  
random samples, I-8  
random sampling  
algorithms, IV-228  
commands, IV-223  
examples, IV-225, IV-226  
overview  
Quick Graphs, IV-224  
univariate continuous, IV-222  
univariate discrete, IV-220  
usage, IV-224  
random variables, II-6  
random walk, IV-517  
randomized block designs, IV-37  
power analysis, IV-19  
range, I-301, I-307, IV-392  
Rank, IV-262  
rank regression, IV-262  
rank-order coefficients, I-172  
Rasch model, IV-490  
receiver operating characteristic curves  
See signal detection analysis  
regression



- bayesian regression, II-50
- LAD regression, IV-260
- Least-squares regression, IV-256
- linear, I-11
- LMS regression, IV-261
- logistic, III-1
- LTS regression, IV-261
- M-regression, IV-261
- rank regression, IV-262
- ridge regression, II-48
- S regression, IV-262
- TSLS regression, IV-581
- two-stage least squares, IV-581
- regression charts, IV-152
- regression trees, I-45
  - algorithms, I-62
  - basic tree model, I-42
  - commands, I-54
  - compared to analysis of variance, I-45
  - compared to stepwise regression, I-46
  - data format, I-54
  - displays, I-51
  - examples, I-55, I-57, I-59
  - loss functions, I-48, I-51
  - missing data, I-62
  - mobiles, I-41
  - model, I-51
  - overview, I-41
  - pruning, I-47
  - Quick Graphs, I-54
  - resampling, I-41
  - saving files, I-54
  - stopping criteria, I-47, I-53
  - usage, I-54
- R-E-G-W Q test, II-197
- R-E-G-W-Q test, II-27, II-119
- reliabilities, IV-492
- reliability, IV-489
- repeated measures, II-31
  - assumptions, II-32
- resampling
  - algorithms, I-38
  - bootstrap-t method, I-19
  - command, I-22
  - examples, I-23, I-27, I-28, I-33, I-34, I-36
  - missing data, I-38
  - naive bootstrap, I-19
  - overview, I-17
  - Quick Graphs, I-22
  - usage, I-22
- response optimization, IV-234
  - canonical analysis, IV-234
  - desirability analysis, IV-236
  - ridge analysis, IV-235
- response surface designs, I-350, I-354
  - analysis of, I-357
  - Box-Behnken designs, I-357
  - central composite designs, I-356
  - examples, I-380, I-384
  - rotatability, I-355, I-356
- response surface methods, IV-231
  - commands, IV-244
  - contour and surface plot, IV-233, IV-243
  - customization, IV-238
  - estimate model, IV-237, IV-238
  - examples, IV-245, IV-247, IV-249, IV-250
  - lack of fit, IV-233
  - optimize, IV-240
  - overview, IV-231
  - Quick Graphs, IV-244
  - usage, IV-244
- response surfaces, I-132, III-273
- restricted/residual maximum likelihood estimates, II-385
- ridge regression, II-48
- right censored data, IV-428
- RMSEA, III-457
- robust discriminant analysis, I-399
- robust regression
  - commands, IV-279
  - examples, IV-280, IV-283, IV-284
  - LAD regression, IV-260
  - LMS regression, IV-261
  - LTS regression, IV-261
  - M-regression, IV-261
  - overview, IV-255

- Quick Graphs, IV-279
- rank regression, IV-262
- S regression, IV-262
- usage, IV-279
- robust smoothing, IV-358, IV-365
- robustness, III-321
- ROC curves, IV-320
- root mean square error of approximation, III-457
- rotatability
  - in response surface designs, I-355
- rotatable designs
  - in response surface designs, I-356
- rotation, I-459
- Roy's Greatest root, III-226
- running median smoothers, IV-512
- running-means smoother, IV-360
- S
  - s charts, IV-126
    - plotting with X-bar charts, IV-129
  - Sakitt D, IV-321
  - sample size, IV-23, IV-30
  - samples, I-8
  - saturated models
    - loglinear modeling, III-95
  - scale regression, IV-262
  - scalogram
    - see partially ordered scalogram analysis with
      - coordinates
  - scatterplot matrix, I-160
  - Scheffé model
    - in mixture designs, I-361
  - Scheffé test, II-27, II-118, II-197, II-307, II-395
  - screening designs, I-360
  - SD-RATIO, IV-321
  - seasonal decomposition, IV-523
  - second-order stationarity, IV-387
  - semi-variograms, IV-388
  - set correlations
    - assumptions, IV-292
    - categorical variables, IV-301
    - data format, IV-304
    - measures of association, IV-293
    - missing data, IV-316
    - overview, IV-291
    - partialing, IV-292
    - usage, IV-304
  - Shapiro-Wilk test, I-302
  - Shepard diagrams, III-189, III-194
  - Shepard's smoother, IV-360
  - Shewhart control charts
    - c charts, IV-131
    - np charts, IV-129
    - p charts, IV-130
    - R charts, IV-128
    - s charts, IV-126
    - u charts, IV-133
    - variance charts, IV-124
    - X charts, IV-129
    - X-bar charts, IV-123
  - Sidak test, II-27, II-118, II-197, II-307, II-395
  - sign test, III-325, III-326
  - signal detection analysis
    - algorithms, IV-346
    - chi-square model, IV-323
    - commands, IV-324
    - convergence, IV-324
    - data format, IV-325
    - examples, IV-328, IV-333, IV-335, IV-336, IV-340, IV-342, IV-344
    - exponential model, IV-323
    - gamma model, IV-323
    - logistic model, IV-323
    - missing data, IV-346
    - nonparametric model, IV-323
    - normal model, IV-323
    - overview, IV-319
    - poisson model, IV-323
    - Quick Graphs, IV-327
    - ROC curves, IV-327
    - usage, IV-325
  - sill, IV-392
  - similarity measures, I-157
  - simple matching dichotomy coefficients, I-164, I-173

- simplex, I-359
- Simplex method, III-269, III-273
- SIMPLS (Straight-forward IMplementation of Partial Least Squares)
  - see partial least squares regression
  - , III-377
- simulation, IV-394
- singular value decomposition, I-201, IV-6, IV-16
- skewness, I-307
  - positive, I-4
- slope, II-13
- smoothing, IV-362, IV-510
  - bandwidth, IV-350, IV-355
  - biweight kernel, IV-362, IV-364, IV-365
  - Cauchy kernel, IV-362, IV-365
  - commands, IV-366
  - confidence intervals, IV-368
  - data format, IV-366
  - discontinuities, IV-360
  - discrete gaussian convolution, IV-361
  - distance-weighted least squares (DWLS), IV-361
  - Epanechnikov kernel, IV-362, IV-364
  - examples, IV-367, IV-368, IV-370, IV-380
  - fixed-bandwidth method, IV-355, IV-362, IV-364
  - Gaussian kernel, IV-362, IV-364, IV-365
  - grid points, IV-361, IV-362, IV-382
  - inverse-distance, IV-360
  - k nearest-neighbors method, IV-356
  - kernel functions, IV-350, IV-352, IV-362, IV-364
  - LOESS smoothing, IV-361, IV-362, IV-367, IV-368, IV-370, IV-380
  - Marron & Nolan canonical kernel width, IV-357, IV-362, IV-364
  - mean smoothing, IV-358, IV-365
  - median smoothing, IV-358
  - methods, IV-350, IV-358, IV-365
  - model, IV-362
  - moving-averages, IV-360
  - Nadaraya-Watson, IV-360
  - nonparametric vs. parametric, IV-350
  - overview, IV-349
  - polynomial smoothing, IV-358, IV-365
  - Quick Graphs, IV-366
  - resampling, IV-349
  - residuals, IV-362, IV-366
  - robust smoothing, IV-358, IV-365
  - running-means, IV-360
  - saving results, IV-364, IV-366, IV-367
  - Shepard's smoother, IV-360
  - step, IV-361
  - tied values, IV-361
  - tricube kernel, IV-364, IV-365
  - trimmed mean smoothing, IV-365
  - triweight kernel, IV-364, IV-365
  - uniform kernel, IV-364
  - usage, IV-366
  - window normalization, IV-357, IV-364
- Sneath and Sokal's binary similarity coefficient, I-164
- Somers' d coefficients, I-227, I-235
- Sorting, I-5
- spaghetti plot, II-458
- spatial statistics, IV-385
  - algorithms, IV-426
  - azimuth, IV-403
  - commands, IV-408
  - data, IV-410
  - dip, IV-403
  - examples, IV-411, IV-417, IV-418, IV-424
  - grid, IV-407
  - kriging, IV-393, IV-400, IV-405
  - lags, IV-402
  - missing data, IV-426
  - model, IV-385, IV-403
  - nested models, IV-392
  - nesting structures, IV-403
  - nugget, IV-392
  - nugget effect, IV-392, IV-405
  - plots, IV-401
  - point statistics, IV-400
  - Quick Graphs, IV-410
  - resampling, IV-385
  - sill, IV-405



- simulation, IV-394, IV-401
- spherical model, IV-404
- trends, IV-406
- usage, IV-410
- variogram, IV-400
- Spearman coefficients, I-162, I-172, I-227
- Spearman-Brown coefficient, IV-489
- specificities, I-458
- spectral models, IV-510
- spherical model, IV-389
- split plot designs, II-175
- split-half reliabilities, IV-492
- SSCP matrix, III-135
- standard deviation, I-3, I-301, I-307
- standard error of estimate, II-7
- standard error of skewness, I-307
- standard error of the mean, I-11, I-307
- standardization, I-67
- standardized alpha, IV-489
- standardized deviates, I-202
- standardized values, I-6
- stationarity, IV-387, IV-520
- statistics
  - defined, I-1
  - descriptive, I-1
  - inferential, I-7
- stem-and-leaf plots, I-3, I-299
- step smoother, IV-361
- stepwise regression, II-15, II-30, III-9
- stochastic processes, IV-386
- stress, III-188, III-211
- structural equation models
  - see path analysis
- Stuart's tau-c coefficients, I-227, I-234
- Student, II-197
- studentized residuals, II-10
- Student-Newman-Keuls test, II-27, II-119
- subpopulations, I-305
- subsampling, I-18
- sum of cross-products matrix, I-171
- sums of squares
  - type I, II-29, II-34, II-113
  - type II, II-35, II-113
  - type III, II-30, II-36, II-113
  - type IV, II-36
- surface plot, IV-243
- surface plots, IV-401
- survival analysis
  - AIC and Schwarz's BIC, IV-427
  - algorithms, IV-476
  - censoring, IV-428, IV-435, IV-479
  - centering, IV-477
  - coding variables, IV-437
  - commands, IV-447
  - convergence, IV-481
  - Cox regression, IV-441
  - data format, IV-448
  - estimation, IV-442
  - examples, IV-449, IV-453, IV-455, IV-459, IV-462, IV-464, IV-468, IV-472
  - exponential model, IV-441
  - graphs, IV-437, IV-444
  - logistic model, IV-441
  - log-likelihood, IV-477
  - lognormal model, IV-435, IV-477
  - missing data, IV-476
  - model, IV-435
  - models, IV-479
  - Nelson-Aalen cumulative hazard estimator, IV-438
  - overview, IV-427
  - parameters, IV-476
  - plots, IV-481
  - proportional hazards models, IV-479
  - Quick Graphs, IV-448
  - Singular Hessian, IV-478
  - stepwise, IV-482
  - stepwise estimation, IV-443
  - tables, IV-437, IV-444
  - time dependent covariates, IV-446
  - usage, IV-448
  - variances, IV-483
  - weibull model, IV-472
- symmetric matrix, I-160



## T

- t tests
- Taguchi designs, I-353, I-377
- Tamhane's T2 test, II-27, II-119, II-197
- Tanimoto dichotomy coefficients, I-164, I-173
- tau-b coefficients, I-234
- tau-c coefficients, I-234
- test for normality, I-302
  - Anderson-Darling test, I-303
  - Shapiro-Wilk test, I-302
- test item analysis
  - algorithms, IV-506
  - classical analysis, IV-488, IV-489, IV-491, IV-506
  - commands, IV-494
  - data format, IV-495
  - examples, IV-498, IV-500, IV-503
  - logistic item-response analysis, IV-490, IV-493, IV-506
  - missing data, IV-507
  - overview, IV-487
  - Quick Graphs, IV-497
  - reliabilities, IV-492
  - resampling, IV-487
  - scoring items, IV-492, IV-493
  - statistics, IV-495
  - usage, IV-495
- tests for correlation, I-535
  - equality of two correlations, I-522, I-537
  - specific correlation, I-522, I-536
  - zero correlation, I-522, I-535
- tests for mean, I-523
  - one-sample t, I-520, I-526
  - one-sample z, I-520, I-523
  - paired t, I-521, I-527
  - poisson, I-520, I-530
  - two-sample t, I-521, I-528
  - two-sample z, I-520, I-524
- tests for normality
  - AD test, III-334
  - K-S test, III-331
  - Lilliefors test, III-334
  - Shapiro-Wilk's test, I-497
- tests for proportion, I-538
  - equality of proportions, I-521
  - equality of two proportions, I-540
  - single proportion, I-520, I-538
- tests for variance, I-531
  - Bartlett's test, I-521
  - equality of several variances, I-534
  - equality of two variances, I-521, I-532
  - Levene's test, I-521
  - single variance, I-531
- tetrachoric correlation, I-164, I-166
- theory of signal detectability (TSD), IV-319
- time domain models, IV-510
- time series, IV-509
  - algorithms, IV-578
  - ARIMA models, IV-514, IV-540
  - clear series, IV-534
  - commands, IV-532, IV-534, IV-539, IV-540, IV-542, IV-544, IV-546
  - data format, IV-546
  - examples, IV-547, IV-548, IV-549, IV-550, IV-552, IV-555, IV-557, IV-558, IV-560, IV-561, IV-566, IV-575
  - forecasts, IV-538
  - Fourier transformations, IV-545
  - missing values, IV-509
  - moving average, IV-511, IV-535
  - overview, IV-509
  - plot labels, IV-528
  - plots, IV-528, IV-529, IV-530, IV-531
  - Quick Graphs, IV-546
  - running means, IV-512, IV-535
  - running medians, IV-512, IV-536
  - seasonal adjustments, IV-523, IV-539
  - smoothing, IV-510, IV-535, IV-536, IV-537
  - stationarity, IV-520
  - transformations, IV-532, IV-534
  - trend analysis, IV-525, IV-542
  - trends, IV-538
  - usage, IV-546
- tolerance, II-16
- T-plots, IV-529

trace criterion

see A-optimality

tree clustering methods, I-47

tree diagrams, I-70

trend analysis, IV-525, IV-542

Homogeneity test, IV-544

Mann-Kendall test, IV-526, IV-543

Modified Seasonal Kendall test, IV-543

Seasonal Kendall test, IV-526, IV-543

slope estimator, IV-573

triangle inequality, III-186

tricube kernel, IV-364

trimmed mean, I-299, I-308

trimmed mean smoothing, IV-365

triweight kernel, IV-364

t-tests, IV-19

one-sample, I-526, IV-50

paired, I-527, IV-51

power analysis, IV-26

two-sample, I-528, IV-53

Tukey procedure, III-279

Tukey test, II-27, II-118, II-196

Tukey's b test, II-27, II-119, II-197

Tukey's HSD test, II-307, II-395

Tukey's jackknife, I-18

twoing, I-48

two-stage least squares

algorithms, IV-597

commands, IV-586

estimation, IV-582

examples, IV-587, IV-590, IV-592, IV-593,

IV-595, IV-596

heteroskedasticity-consistent standard errors,

IV-586

lagged variables, IV-586

missing data, IV-597

model, IV-585

overview, IV-581

Quick Graphs, IV-586

usage, IV-586

Type I error, IV-21

Type II error, IV-22

## U

u charts, IV-133, IV-134

unbalanced designs

in analysis of variance, II-29

uncertainty coefficient, I-234

unfolding models, IV-3

uniform kernel, IV-364

## V

validity, I-87

variance, I-307

of estimates, I-355

variance charts, IV-124

variance component models

see mixed regression

variance components

categorical variables, II-303

commands, II-310

examples, II-311, II-315, II-320, II-323, II-326, II-328, II-334, II-340

hypothesis test, II-306

model estimation, II-301

models, II-301

options, II-304

overview, II-299

Quick Graph, II-310

usage, II-310

variance inflation factor, II-70

variance of prediction, I-356

variance paths

path analysis, III-401

varimax rotation, I-460, I-464

variograms, IV-388, IV-401

model, IV-389

vector model

in perceptual mapping, IV-5

Voronoi polygons, IV-385, IV-397, IV-400

## W

Wald-Wolfowitz runs test, III-337

wave model, IV-391

Weibull, III-334  
 Weibull distribution, IV-432  
 weighted running smoothing, IV-512  
 weights, I-23, I-54, I-135, I-179, I-206, I-246, I-248, I-323, I-371, I-408, I-469, I-503, I-544, II-54, II-121, II-122, II-202, II-311, II-357, II-399, II-441, II-442, III-23, III-103, III-104, III-137, III-194, III-217, III-283, III-339, III-340, III-364, III-385, III-413, IV-9, IV-63, IV-104, IV-162, IV-244, IV-280, IV-305, IV-325, IV-328, IV-366, IV-367, IV-410, IV-449, IV-495, IV-498, IV-547, IV-587

Wilcoxon Signed-Rank test, III-326

Wilcoxon test, III-326

Wilk's trace, I-405

Wilks' lambda, I-405, III-225

Winter's three-parameter model, IV-524

Within-Group Testing, III-241, III-257

within-subjects differences

in analysis of variance, II-32

## X

X charts, IV-129

X-bar charts, IV-123

plotting with R charts, IV-129

plotting with s charts, IV-129

X-MR charts, IV-149

control limits, IV-149

## Y

Yates' correction, I-226, I-233

y-intercept, II-12

Young's S-STRESS, III-190

Yule's Q, I-228

Yule's Q coefficient, I-164

Yule's Y, I-228, I-234

## Z

z tests

z-tests, IV-19

one-sample, IV-46

two-sample, IV-48